

the
compositionality
papers

compositionality
papers

Jerry A. Fodor and Ernest Lepore

OXFORD

Copyrighted material

THE COMPOSITIONALITY PAPERS

This One



BT8S-5DR-TGKP

The Compositionality Papers

JERRY FODOR
and
ERNEST LEPORE

CLARENDON PRESS · OXFORD

OXFORD

UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Bangkok Buenos Aires Cape Town Chennai
Dar es Salaam Delhi Hong Kong Istanbul Karachi Kolkata
Kuala Lumpur Madrid Melbourne Mexico City Mumbai Nairobi
São Paulo Shanghai Singapore Taipei Tokyo Toronto
and an associated company in Berlin

Oxford is a registered trade mark of Oxford University Press
in the UK and certain other countries

Published in the United States
by Oxford University Press Inc., New York

© in this volume Jerry A. Fodor and Ernest Lepore, 2002

The moral rights of the authors have been asserted
Database right Oxford University Press (maker)

First published 2002

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system, or transmitted, in any form or by any means,
without the prior permission in writing of Oxford University Press,
or as expressly permitted by law, or under terms agreed with the appropriate
reprographics rights organizations. Enquiries concerning reproduction
outside the scope of the above should be sent to the Rights Department,
Oxford University Press, at the address above

You must not circulate this book in any other binding or cover
and you must impose the same condition on any acquirer

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Fodor, Jerry A.

The compositionality papers / Jerry Fodor and Ernest Lepore.
p. cm.

Rev. papers, previously published in various sources, 1991–2001.

Includes bibliographical references and indexes.

Contents: Why meaning probably isn't conceptual role—The pet fish and the red
herring: why concepts aren't prototypes—Why compositionality won't go away:
reflections on Horwich's "deflationary" theory—What can't be valued can't be valued,
and it can't be supervalued either—The emptiness of the lexicon: reflections on
Pustejovsky—Impossible words?—Brandom's burdens: compositionality and
inferentialism—Churchland and state space semantics—All at sea in semantic space:
Churchland on meaning similarity.

1. Language and languages—Philosophy. I. Lepore, Ernest, 1950– II. Title.

P107 .F63 2002 401—dc21 2002019756

ISBN 0-19-925215-7 (hbk)

ISBN 0-19-925216-5 (pbk)

1 3 5 7 9 10 8 6 4 2

Typeset by SNP Best-set Typesetter Ltd., Hong Kong
Printed in Great Britain by
T.J. International Ltd, Padstow, Cornwall

ACKNOWLEDGMENTS

Thanks to the original publishers for permission to reproduce the following.

'Why Meaning (Probably) Isn't Conceptual Role' was originally published in *Mind and Language*, 6/4 (1991), 329–43. It is reprinted here with the permission of Basil Blackwell.

'The Pet Fish and the Red Herring: Why Concepts Aren't Prototypes' was originally published in *Cognition*, 58/2 (Feb. 1996), 243–76. It is reprinted here with the permission of *Cognition*.

'Why Compositionality Won't Go Away: Reflections on Horwich's "Deflationary" Theory' was originally published in *Ratio*, 14/4 (2001), 350–68. It is reprinted here with the permission of the publishers.

'What Can't Be Valued Can't Be Valued, and it Can't Be Supervalued Either' was originally published in the *Journal of Philosophy*, 93 (1996), 516–36. It is reprinted here with the permission of the publishers.

'The Emptiness of the Lexicon: Reflections on Pustejovsky' was originally published in *Linguistic Inquiry*, 29/2 (Spring 1998), 269–88. It is reprinted here with the permission of the publishers.

'Impossible Words?' was originally published in *Linguistic Inquiry*, 30/3 (Summer 1999), 445–53. It is reprinted here with the permission of the publishers.

'Brandom's Burdens: Compositionality and Inferentialism' was originally published in *Philosophy and Phenomenological Research*, 73/2 (Sept. 2001), 465–81. It is reprinted here with the permission of the publishers.

'Paul Churchland and State Space Semantics' was published in R. McCauley (ed.), *The Churchlands and their Critics* (Basil Blackwell, 1996), 145–58. It is reprinted here with the permission of the publishers.

'All at Sea in Semantic Space: Churchland on Meaning Similarity' was originally published in the *Journal of Philosophy*, 96/8 (Aug. 1999), 381–403. It is reprinted here with the permission of the publishers.

All the papers underwent at least stylistic revisions; and there are substantive alterations, of one kind or another, in 'The Pet Fish and the Red Herring: Why Concepts Aren't Prototypes' and 'The Emptiness of the Lexicon'. In addition, we added appendices to both 'What Can't Be Valued Can't Be Valued, and it Can't Be Supervalued Either' and 'All at Sea in Semantic Space: Churchland on Meaning Similarity'.

Thanks to Sarah-Jane Leslie for all her help in collecting the papers. We would also like to thank Peter Momtchiloff, Charlotte Jenkins, and Laurien Berkeley of Oxford University Press.

CONTENTS

Geographical Introduction	1
---------------------------	---

Part I. Composition

1. Why Meaning (Probably) Isn't Conceptual Role	9
2. The Pet Fish and the Red Herring: Why Concepts Still Can't Be Prototypes	27
3. Why Compositionality Won't Go Away: Reflections on Horwich's 'Deflationary' Theory	43
4. What Can't Be Valued Can't Be Valued, and it Can't Be Supervalued Either	63

Part II. Decomposition

5. The Emptiness of the Lexicon: Reflections on Pustejovsky	89
6. Impossible Words?	120

Part III. New Pragmatists

7. Brandom's Burdens: Compositionality and Inferentialism	135
8. Paul Churchland and State Space Semantics	155
9. All at Sea in Semantic Space: Churchland on Meaning Similarity	174

<i>References</i>	201
-------------------	-----

<i>Name Index</i>	207
-------------------	-----

<i>Subject Index</i>	209
----------------------	-----

Geographical Introduction

COMPOSITIONALITY is the property that a system of representation has when (i) it contains both primitive symbols and symbols that are syntactically and semantically complex; and (ii) the latter inherit their syntactic/semantic properties from the former. For example, it is part and parcel of the compositionality of English that the symbol 'dogs' is complex, that it has the symbols 'dog' and 's' as constituents, and that its meaning (viz., *dogs*) is inherited from the meanings of 'dog' (viz., *dog*) and 's' (viz., *plural*). Likewise, it is part and parcel of the compositionality of English that the symbol 'John jumps' is complex, that it has among its constituents the symbols 'John' and 'jumps', and that its meaning (viz., *John jumps*) is inherited from the meanings of these subsentential parts. The details are moot, but the general intuition will suffice for our present purposes: It seems clear (and we assume in what follows) that all the systems of representation that philosophers/cognitive scientists are professionally concerned with are more or less compositional. In particular, natural languages are, and so too is the system of mental representations that presumably mediates such mental processes as perception and thought.

Briefly, the standard arguments for this claim are that language and thought are both productive and systematic. Productivity is the property that a system of representation has when it contains an infinite number of syntactically and semantically distinct symbols (as, for example, English contains the open-ended sequence of nonsynonymous expressions: 'missile shield', 'anti-missile shield',

'anti-anti-missile-shield-shield', 'anti-anti-anti-missile-shield-shield-shield', and so forth until President Bush bankrupts us all). Systematicity is the property that a system of representation has when it contains such families of syntactically and semantically related but distinct expressions as: 'John loves Mary', 'Mary loves John', 'Does John love Mary?', 'Does Mary love John?', and so forth.

It appears not to be seriously in doubt that the compositionality of mental and linguistic representation underlies both their productivity and their systematicity. Mental and linguistic representations are constituted by a finite number of recurring primitive parts, the arrangement of which determines the structure and content of all the complex representations. Because the primitive parts can occur in indefinitely many configurations, new complex symbols can be produced without bound. Likewise, the systematicity of a family of expressions consists in the partial overlap of their primitive parts. Despite a substantial literature to the contrary (e.g., on the part of connectionists), neither the facts of productivity and systematicity, nor their essential connection to compositionality, are subject to serious doubt. There are familiar technical problems about explaining how compositionality works in representational systems that exhibit (for example) vagueness intensional contexts, nondenoting terms, empty predicates, indexicals, token reflexives and other contextually sensitive expressions, idioms, etc. The present collection of papers considers some of these, but by and large we're content to leave them to the experts. Our main concern is with the connections between compositionality and two major galaxies of problems about language and mind.

Compositionality and the Structure of Concepts

Questions about what concepts are, and what it is for a mind to have one, predominate in centuries of psychology and the philosophy of mind (and in linguistics too, insofar as word meanings are widely identified with concepts). It makes a lot of difference how these questions are answered. Suppose, for example, that *having a concept* is something like *having a theory about whatever it is that the concept applies to*. Presumably it follows that concept acquisition must be a

lot like theory construction; both are species of nondemonstrative, abductive inference. Correspondingly, infants must be much like scientists, only smaller. Psychologists have actually held that view (Gopnik and Meltzoff 1997; Karmiloff-Smith 1992; etc.). Or suppose you think that concepts are something like summaries of the typical properties of the things that they apply to (for example, that they are sets of weighted 'semantic features'). Presumably it follows that concept acquisition is largely a process of nonabductive, nondemonstrative inference; specifically, that it's a species of statistical inference. Psychologists have actually held that view too. (See, e.g., Elman *et al.* 1996.)

So it matters to psychology and semantics what concepts are. It also matters to epistemology. Suppose you hold a kind of operationalism, according to which a concept is something like a sorting procedure and having a concept is having a method for determining whether a given object falls in its extension. It presumably follows that to have the concept DOG is to know how to tell, at least in favorable circumstances, that something is a dog. But now, philosophical skeptics deny that there are *any* circumstances, however favorable, in which one can tell whether something is a dog. So then, skepticism about dogs is refuted by a proper understanding of what it is to have the concept DOG. That's what epistemologists call a 'paradigm case' argument. There actually are some epistemologists who hold that such arguments are sound.

But although it matters a lot, in a lot of respects, what concepts are, consensus on this issue has not been easy to come by. Here are some familiar candidates. Concepts are: mental images; sorting procedures; licenses to perform inferences; word meanings; abstractions from language games; abstractions from theories; patterns of deference; names of senses; names of denotations; stereotypes; prototypes; definitions. Whatever. This is an embarrassment of riches, among which it has proved remarkably hard to choose.

We think, however, that compositionality chooses; that it does, and why it does, are main themes in the papers that follow. The line of our argument is that since mental representation and linguistic meaning are *de facto* compositional, we can reject out of hand any theory that says that concepts(/word meanings) are Xs unless Xs are the sorts of things for which compositionality holds. That is, there

must be a distinction between primitive and complex Xs, and the syntactic/semantic properties of the latter must be inherited from the syntactic/semantic properties of the former. Otherwise, *the proposed identification of concepts(/word meanings) with Xs is ipso facto refuted, however attractive it may otherwise appear*. It's our belief that, of the various familiar candidates for concepthood, only very few—perhaps only one—can meet this compositionality condition; it's the bull in almost everybody's china shop. If that's right, then, compositionality tells us what concepts and word meanings are.

If many concepts are semantically and structurally complex, then some concepts must be semantically and structurally primitive; likewise word meanings *mutatis mutandis*. A theory of compositionality is thus concerned, willy-nilly, with which concepts are which. There is a view about this that is markedly predominant in cognitive science, namely that most everyday concepts (roughly, the ones that are expressed by high-frequency content words like 'dog', 'chair', 'tree', 'face', and the like) are structured; perhaps they are sets of 'semantic features', so that, for example, DOG is a complex concept whose constituents are such other concepts as IS AN ANIMAL, TYPICALLY BARKS, TYPICALLY HAS A TAIL, or whatever. Lots of philosophers have likewise held that sentences that express nonlogical necessities ('dogs are animals', 'bachelors are unmarried', and so forth) are *ipso facto* analytic, and that analytic sentences are true in virtue of *containment relations among the concepts they express*. It's necessary that dogs are animals because (the feature) ANIMAL is literally a part (a constituent) of the (complex) concept DOG. So, just as issues about the nature of compositionality connect with questions about which concepts are primitive, so too do questions about which concepts are primitive connect with issues about the nature of conceptual necessity.

We are inclined to be atomists about everyday concepts. We think that the concept DOG is probably primitive, and that because it is, it is possible to have the concept DOG without having (e.g.) the concepts ANIMAL or BARKS. These views of ours are widely regarded as eccentric, and we don't argue for them here (but see Fodor and Lepore 1992). We do, however, examine a variety of examples from the recent literature in 'linguistic semantics' that have been supposed to make a case that some morphemically simple words express con-

cepts of considerable structural complexity. Perhaps there are some interesting examples of sentences that are necessary because they are analytic, and perhaps there are some interesting cases where these analyticities derive from relations of conceptual containment. But we're going to argue that the evidence for any of that is negligible to nil.

It's widely agreed among historians of philosophy that sometime around 1950 (or maybe even earlier) the discipline took a 'linguistic turn'. Whereas it used to be thought (by Hume and Kant, for example) that epistemology would underwrite a critical method in philosophy ('when we run over libraries persuaded of these principles, what havoc must we make?'), it became the new working consensus that semantics would do so instead. The idea was to run a sort of transcendental argument based on the prerequisites for meaning. For example: Linguistic(/conceptual) content has to be learnable (or it has to be teachable; or it has to be public; or it has to be social; or it has to be accessible from the perspective of a Radical Translator; or it has to be accessible from the perspective of an Ideal Interpreter; or it has to be a kind of Knowing How rather than a kind of Knowing That, etc., etc.). These prerequisites for content would, in turn, constrain what concepts and word meanings(/the grasp of concepts and word meanings) could possibly be. So, for example, there's a transcendental argument that runs: 'Concepts/word meanings can't be introspectible mental images. For, if they were, then word meanings wouldn't be public (or teachable; or learnable; or accessible to Radical Translators) since nobody but you can introspect your images.'

In the event, the linguistic turn didn't lead anywhere; or worse, it led to pragmatism. That wasn't, we think, because the idea of a semantic, transcendental deduction of interesting properties of linguistic/conceptual content is inherently a bad one. Rather, it's because the premises that such deductions assumed were generally empiricist, hence unsound. For example, it begs the question against nativism to suppose that meanings have to be learnable, or teachable, or public, or Radically Translatable, or whatever; the content of mental representations may well be a counterexample of all of these theses. That being so, it's unsurprising that transcendental arguments that assume the contrary don't convince.

But we think there is, after all, a pretty convincing argument of the general form that the linguistic turn imagined: It runs from the premise that linguistic and conceptual meanings are compositional to some quite striking conclusions about what kind of things linguistic and conceptual meanings must be. We think you can get from such arguments considerable insight into the nature of content, hence into the nature of language and mind. So, anyhow, the papers in this collection undertake to argue. So, anyhow, we hope they will persuade you.

Part I

Composition

Why Meaning (Probably) Isn't Conceptual Role

IT'S an achievement of the last couple of decades that people who work in linguistic semantics and people who work in the philosophy of language have arrived at a friendly, *de facto* agreement as to their respective job descriptions. The terms of this agreement are that the semanticists do the work and the philosophers do the worrying. The semanticists try to construct actual theories of meaning (truth theories, or model theories, or whatever) for one or another kind of expression in one or another natural language; for example, they try to figure out how the temperature could be rising compatibly with the substitutivity of identicals. The philosophers, by contrast, keep an eye on the large, foundational issues, such as: What's the relation between sense and denotation; what's the relation between thought and language; whether translation is determinate; and whether life is like a fountain. Every now and then the philosophers and the semanticists are supposed to get together and compare notes on their respective progress. Or lack thereof.

Accordingly, this paper is about not semantics but the philosophy of language. Some of us have been poking around in the basement of meaning theory, and we seem to have discovered a large, nasty crack; as far as we can tell, one of the foundation stones is coming unstuck. We thought we'd better tell you about it before the thing gets worse.

We'll proceed as follows: First we'll try to say where in the foundations the problem is located; then we'll try to say what the problem is; and then we'll make a suggestion or two about what to do about the problem. The first part of the discussion will be very broad; the

second part will be rather more specific; the third part will be practically nonexistent.

Here goes.

Where the Problem Is

A traditional foundational issue in the theory of meaning is: Where do semantic properties come from? The presupposition of this question is that the fact that a word (or a sentence, or whatever) means what it does can't be a brute fact. It can't be a brute fact, for example, that 'dog' means *dog* and not *proton* and that 'proton' means *proton* and not *dog*. Rather, 'dog' must have some nonsemantic property in virtue of which it means *dog* and not *proton*; and 'proton' must have some (different) nonsemantic property in virtue of which it means *proton* and not *dog*. To put it in the standard philosophical jargon, semantic properties must supervene on nonsemantic properties. There may be some properties that things just have; that they have for no reason at all. But if there are, they are presumably the kinds of properties that basic physics talks about (like mass, charge, and charm). They certainly don't include the kinds of properties that semanticists talk about (like meaning *dog* or being a synonym of 'bachelor').

We remark in passing that none of this is to be construed as an attempt to legislate physicalism. For present purposes we're content that semantic properties should be, for example, irreducibly intentional, or irreducibly epistemological, or irreducibly teleological. But we take it not to be in the cards that they are irreducibly semantic. In short, we don't care whether semantic properties supervene on something that is physical just as long as they supervene on something other than themselves.

So the question arises: What do the semantic properties of symbols supervene on? Over the years, philosophers of language have been enthusiastic for two quite different (perhaps incompatible) kinds of answers to this question. There's what we will call the 'Old Testament' story, according to which the meaning of an expression supervenes on the expression's relation to things in the world; and there's what we will call the 'New Testament' story, according to which the

meaning of an expression supervenes on the expression's role in a language. The disagreement between these two sorts of stories is venerable, and we don't propose to go into the details here. Just a paragraph or two by way of assembling reminders.

Old Testament Semantics derives most directly from the British Empiricists and has, among its modern representatives, behaviorist psychologists like Watson and Skinner and a handful of philosophers in the 'naturalized semantics' movement (including Dretske, Millikan, Barwise, and Perry, and, in certain of their moods, one or more of the Fodors). The basic idea is that 'dog' means *dog* because of some (nonsemantic) relation that holds between the symbol and the animal. If you are a British Empiricist (or, at least, if you're Hume), you say that this relation comes down to some variant on resemblance; 'dog' means *dog* rather than *proton* because 'dog' is associated with a certain mental image that means *dog* rather than *proton*. And the mental image means *dog* rather than *proton* because it resembles impressions of dogs quite a lot but resembles impressions of protons hardly at all. If you are a behaviorist psychologist, you say that the symbol-world relation that semantic properties supervene on is *causal* (typically associative); 'dog' means *dog* rather than *proton* because, in consequence of the speaker's history of conditioning (or something), dogs cause utterances of 'dog' and protons don't. Recent developments of Old Testament Semantics propose still other variations on this theme, including, for example, appeals to nomological and/or informational relations between symbols and the world.

This Old Testament story about meaning has come under a lot of criticism, both in philosophy and in linguistics. Perhaps the basic objection is one that derives from Frege: Meaning *can't* be a symbol-world relation, according to this objection, because *identity* of symbol-world relations is compatible with *difference* of meaning. Thus, it's plausible that the expressions 'the Morning Star' and 'the Evening Star' are both attached to the same nonlinguistic thing (viz., to Venus); but it's also plausible that they don't *mean* the same thing. If they did, you couldn't deny that the Morning Star is the Evening Star without self-contradiction or assert that it is without tautology. Both of which, in fact, you can.

So, then, according to this argument, the meaning of an expression

doesn't supervene on the way that it's attached to the world. So what *does* it supervene on?

The New Testament story is an elaboration of the following idea: the expressions 'the Morning Star' and 'the Evening Star' mean different things, despite their both being attached to Venus, because they have different *roles in the (English) language*. 'Only in the context of a sentence does a word have a meaning,' says Frege, and Wittgenstein adds that 'to understand a sentence is to understand a language'. The core idea is that it's the way they are connected *to one another* that determines what the expressions in a language mean. Very often (in fact, in all the semantic theories that we will discuss) this notion of 'role in the language' is given an epistemological twist: To master a linguistic expression is to be disposed to draw some core set of *inferences* that fix its semantic relation to other expressions in the language. To master the expression 'dog' is, *inter alia*, to be prepared to infer from 'Rover is a dog' to 'Rover is an animal'. To master the expression 'the Morning Star' is, *inter alia*, to be prepared to infer from 'is the Morning Star' to 'rises in the morning'. And so forth.

We want to emphasize that, at this stage, we're using notions like *infer* and *disposed to infer* as blank checks. For the moment we'll let any disposition to have one belief cause another count as a disposition to infer the second from the first. So deductive inferences, and inductive inferences, and plausible inferences, and prudential inferences, and mere associations, and heaven knows what all else, are included *pro tem*. A lot of this paper will be about problems that arise when a New Testament Semanticist tries to say exactly which inferences are constitutive of the meanings of the terms that enter into them.

The New Testament picture in philosophy is quite close to one that's familiar from structuralist linguistics, according to which the meaning of an expression is its role in a 'system of differences'. To know what 'dog' means is to know that it excludes 'cat' and cohabits with 'animal'. To know what 'bachelor' means is to know that it excludes 'spinster' and cohabits with 'unmarried'. Notice that it follows from this view that if English didn't contain words that mean what 'cat' and 'animal' do, *it couldn't contain a word that means what 'dog' does*. This is a kind of conclusion that structuralists in seman-

tics have, with varying degrees of enthusiasm, quite generally been prepared to embrace.

Although, as far as we can tell, one or another version of the New Testament view is held practically as gospel not only in linguistics and philosophy, but throughout the cognitive sciences, it is nevertheless possessed of well-known problems. For example, New Testament Semantics appears to be intractably holistic. This is because, once you start identifying the content of a belief with its inferential role in a system of beliefs (*mutatis mutandis*, the meaning of an expression with its inferential role in a language), it's hard to avoid proceeding to identify the content of a belief with its *whole* inferential role in the system of beliefs. And, having gone that far, it's hard to avoid the conclusion that if belief systems differ at all with respect to the propositions they endorse, then they differ completely with respect to the propositions they endorse. (*Mutatis mutandis*, if languages differ at all in respect of the propositions they can express, then they differ entirely in respect of the propositions they can express.¹) This is a well-greased, and well-traveled, slippery slope; having arrived at the bottom, one finds oneself accepting such *prima facie* outlandish doctrines as that no two people ever share a belief; that there is no such relation as translation; that no two people ever mean the same thing by what they say; that no two time slices of *the same* person ever mean the same thing by what they say; that no one can ever change his mind; that no statements, or beliefs, can ever be contradicted (to say nothing of refuted); and so forth. It's a moot question how to get the goodness out of inferential role semantics without paying this extravagant price. Indeed, it's moot whether it is possible to do so.

Serious though these worries are, however, they are all, as it were, external to semantics as such; they impugn its consequences for epistemology and ontology rather than its actual coherence. And a convinced New Testament Semanticist might be prepared to bite the bullet. If the doctrine that meaning supervenes on intralinguistic relations has relativistic, idealistic, or even solipsistic consequences,

¹ The issues we will be concerned with arise in precisely parallel ways for theories of the meanings of linguistic expressions and for theories of the contents of thoughts. In the rest of what follows, we will not bother to draw this distinction unless something turns on it.

then perhaps we had better learn to live with relativism, idealism, or even solipsism. There are many, especially on the West Coast, who clearly long to do so.

Well, so much for the background; now here's the problem. We want to argue that, quite aside from ontology and epistemology, it looks like New Testament Semantics can't be squared with a pair of linguistic hypotheses that practically everybody thinks there are very good reasons to endorse; and which, *de facto*, most New Testament Semanticists actually do endorse. These are, first, that natural languages are compositional, and, second, that the analytic/synthetic distinction is unprincipled. (This means not that the distinction is *vague*—what empirical distinction isn't?—but that there aren't any expressions that are true or false solely in virtue of what they mean. Or, if there are, that nobody can say without begging questions which ones.) The foundational problem to which we wish to call your attention is that these two principles together with New Testament Semantics form an inconsistent triad. At least one of the three will have to go.

The next section of the paper sketches our argument for this claim; the third asks what on earth to do about it.

Compositionality, Analyticity, and Inferential Role Semantics

There is, we believe, an internal connection between analyticity and compositionality that hasn't previously been remarked upon and that has dire implications for the New Testament identification of meanings with inferential roles. We commence our exploration of this territory with a few remarks about compositionality.

A language is *compositional* if and only if (idioms aside) the meaning of its complex expressions is inherited from the meaning of their syntactic structures together with the meaning of their syntactic constituents. We take the doctrine that natural languages are compositional to be, as one says in Britain, nonnegotiable.

No doubt connectionist psychologists (and a handful of philosophers) have occasionally been moved to dispense with compositionality. But that just indicates the desperateness of their plight. For compositionality is at the heart of some of the most striking proper-

ties that natural languages exhibit. The most obvious of these are *productivity* (roughly, the fact that every natural language can express an open-ended set of propositions) and *systematicity* (roughly, the fact that any natural language that can express the proposition P will also be able to express many propositions that are semantically close to P . If, for example, a language can express the proposition that aRb , then it can express the proposition that bRa ; if it can express the proposition that $P \rightarrow Q$, then it can express the proposition that $Q \rightarrow P$; and so forth).²

Closely connected to productivity and systematicity is a feature of natural languages that is so patent that it's easy to ignore. Consider the indefinite range of expressions ' Fa ', ' Ga ', ' Ha ', and so forth. It's self-evident, on the one hand, that the individual that ' a ' refers to is implicated in all of their truth conditions (e.g., ' Fa ' is true iff a is in the extension of ' F '; ' Ga ' is true iff a is in the extension of ' G ', etc.). What explains such correlations between the semantic properties of sentences and their syntactic constituency? Pretty clearly, it's that ' a ' contributes *the same semantic value* (viz., the individual a) to each of the expressions it occurs in. In similar spirit, it is notable that ' a ' is the *only* singular term in ' Fa ' and that, on the one hand, a is the only individual that ' Fa ' assigns to the extension of ' F ', and, on the other hand, F is the only extension to which ' Fa ' assigns a . What explains all this is presumably that only what ' a ' and ' F ' contribute can belong to the truth conditions of ' Fa '. Collecting all these truisms, we get the following: A given syntactic expression contributes the same semantic content to each of its complex hosts; and each host expression receives its semantic content from its syntactic constituents. But a language that satisfies these principles is compositional by definition.

Our point is that the various generalizations we've just been considering are connected and explained on the assumption of compositionality; they are all consequences of the principle that the meaning

² It's sometimes suggested that pairs like 'John calculated the answer' but *'the answer calculated John' show that English is less than entirely systematic, hence that the argument from systematicity to compositionality is unsound. However, even an imperfect systematicity is more than adequate for the present polemical purposes. There is a vast number of cases where $aRbs$ have corresponding $bRas$, and, more generally, where aXs have corresponding aYs . All such cases must be accidental unless compositionality is assumed to be the default.

nevertheless serve as tools in the interpretation of other, binding, human rights instruments.

Turning to human rights treaties, there are no substantive provisions indicating an intention to impose obligations directly on individuals in the International Covenant on Economic, Social and Cultural Rights. Some have seen in Article 5(1) of the Civil and Political Rights Covenant and Article 17 of the European Convention the imposition of individual obligations.³¹ The provisions are basically interpretation clauses, indicating that the treaties should not be construed as justifying conduct trenching upon or denying human rights guaranteed by the treaty. The absence of a right does not signal the presence of an obligation, however, and there is nothing in these provisions indicating the imposition of an individual duty not to interfere with the rights of others.³² Article 19(3) of the Civil and Political Rights Covenant and Article 10(2) of the European Convention provide that the exercise of freedom of expression 'carries with it special duties and responsibilities', so that legal restriction may lawfully be imposed by the state. The nature and content of such duties are not articulated, and they seem to have been included essentially as justification for the 'clawback' clause permitting limitation of the right.³³ Nevertheless, they do represent an acknowledgment that human duties may be envisaged within the framework of those two instruments.

The American Convention on Human Rights for its part contains a chapter on 'personal responsibilities'. Its sole article, entitled 'Relationship between duties and rights' (Art. 32), provides very broadly that '[e]very person has responsibilities to his family, his community, and mankind' (para. 1), and that the rights of individuals under the Convention are limited by a number of factors, including the rights of others (para. 2). While the second paragraph imposes no obligation, being merely concerned with the limitation of rights, the first recognises the existence of individual duties, including duties towards other individuals.³⁴ This acknowledgment, however, is made at a most general

³¹ Meron, *Human Rights in Internal Strife*, at 34–5; Jordan Paust, 'The Other Side of Rights: Private Duties Under Human Rights Law', (1992) 5 *Harv. Hum. Rts J* 51, 55.

³² This aspect of Art. 5(1) is discussed by the UN Human Rights Committee in *López v. Uruguay*, Communication No. 52/1979, reprinted in 'Human Rights Committee – Selected Decisions Under the Optional Protocol', UN Doc. CCPR/C/OP/1 (1985) (Sales No. E.84.XIV.2), at 91.

³³ See e.g. *Handyside v. United Kingdom*, (1968) Ser. A No. 24, para. 49 (Eur. Ct Hum. Rts).

³⁴ Héctor Gros Espiell, *La Convención americana y la Convención europea de derechos humanos – Analysis comparativo* (Santiago: Ed. jurídica de Chile, 1991) 133–5. Unlike the

'dangerous', so, to this extent, it doesn't look like the inference from 'brown cow' to 'dangerous' is compositional.

In short, it appears that some, but not all, of the inferential potential of 'brown cow' (some of its role in one's language or belief system) is determined by the respective inferential potentials of 'brown' and 'cow', the rest being determined by one's 'real world' beliefs about brown cows. This should not seem surprising or contentious; it's just a way of saying that 'brown cows are dangerous' (unlike 'brown cows are animals' and 'brown cows are not-green' and 'brown cows are brown') is clearly synthetic. That is 'brown cows are dangerous' is contingently true, true in virtue of facts about brown cows, not just in virtue of facts about meanings (assuming that it's true at all).

But, to repeat, if meanings are compositional and inferential roles aren't, then it follows that meanings can't be inferential roles.

As far as we can tell, this line of argument is quite robust; in particular, it doesn't depend on detailed assumptions about how semantics construes the notion of inferential role. There is, for example, an influential paper by Hartry Field in which inferential role is analyzed in terms of subjective probabilities; in effect, the inferential role of your thought that P is identified with the subjective probability that you (would) assign to P contingent on each of the other thoughts that you (can) entertain (Field 1977). So, for example, the inferential role of your thought that it's raining is determined in part by the subjective probability that you (would) assign to that thought on the assumption that the streets are wet; and in part by the subjective probability that you (would) assign to it on the assumption that the sun is out; and in part by the subjective probability that you (would) assign to it on the assumption that elephants have wings . . . and so forth. Like other species of New Testament Semantics, this treatment generalizes, in fairly obvious ways, from a theory according to which inferential roles are assigned to beliefs to one in which they are assigned to linguistic expressions, or to both.

Our point is that the construal of inferential roles in terms of subjective probabilities, whatever other virtues it may have, does nothing to help with the compositionality problem. This is because *subjective probabilities are not themselves compositional*. For example, the subjective probability one assigns to the thought that *brown cows are dangerous* is not a function of the subjective probability one assigns to

the thought that *cows are dangerous*, together with the subjective probability that one assigns to the thought that *brown things are dangerous*. If this seems not obvious, consider a world (or rather a belief world, since the probabilities at issue are supposed to be subjective) where there are very many things that are cows, almost none of which is dangerous, and very many things that are brown, almost none of which is dangerous, and a very small number of brown cows, almost all of which are very, very fierce. On these assumptions, the probability that something that is brown is dangerous is small, and the probability that something that is a cow is dangerous is small, but the probability that a brown cow is dangerous is as big as you please.

We intend the argument so far as one horn of a dilemma, and we anticipate the following reply: "OK, so if the compositionality of meaning is assumed, meanings can't be identified with inferential roles as such. But this doesn't really embarrass New Testament Semantics because meanings can still be identified with roles in *analytic* inferences. Thus, on the one hand, the inference 'brown cow' $\rightarrow$ 'brown animal' is compositional (it's inherited from the inference 'cow' $\rightarrow$ 'animal') and, on the other hand, precisely because it is compositional, 'brown cow' $\rightarrow$ 'brown animal' is analytic. Compositional inferences will always be analytic and analytic inferences will always be compositional; the compositionality of an inference and its analyticity are the same thing".

Look at it this way: If 'brown cow' $\rightarrow$ 'brown animal' is compositional, then it's warranted by the inferential roles of the expressions 'brown' and 'cow'. That's what it is for an inference to be compositional. But, according to New Testament Semantics, the inferential roles of 'brown' and 'cow' *are their meanings*. So, then, that 'brown cow' $\rightarrow$ 'brown animal' is warranted follows from the *meanings* of 'brown' and 'cow'. But for an inference to be analytic *just is* for it to be warranted by the meanings of its constituent expressions. So the compositionality of 'brown cow' $\rightarrow$ 'brown animal'—or, *mutatis mutandis*, of any other inference—*entails* its analyticity. The same argument also works the other way around. For an inference to be analytic is for it to be warranted by the meanings of its constituents. But, according to New Testament Semantics, meanings are inferen-

tial roles. So, for an inference to be analytic is for it to be warranted by the inferential roles of its constituents. But for the warrant of an inference to be determined by the inferential roles of its constituents is for the inference to be compositional. So compositionality entails analyticity and vice versa. So, then, meaning is compositional and role in analytic inference is compositional but inferential role per se is not compositional. What all that shows is just that *we need a revised version of New Testament Semantics—one which identifies meaning with role in analytic inference.*

The first thing to say about this new suggestion is that the threat of circularity is now very close to the surface. It is proposed that we reconcile New Testament Semantics with the compositionality of meaning by identifying the meaning of an expression with its role in analytic inferences. But the difference between analytic inferences and inferences *tout court* is just that the validity of the former is guaranteed *by the meanings* of their constituent expressions. So analyticity and meaning (and compositionality) scrape out a living by doing one another's wash, and Quine gets to say 'I told you so'.

We can now say pretty exactly what our problem is: You can't identify meanings with inferential roles *tout court*, since, unlike meanings, inferential roles *tout court* aren't compositional. You can identify meanings with roles in analytic inferences, however, because analytic inferences are compositional. But, of course, the cost of identifying meanings with roles in analytic inferences is buying into the analytic/synthetic distinction. So the cost of New Testament Semantics is buying into the analytic/synthetic distinction. But, these days, practically everybody thinks that the analytic/synthetic distinction is unprincipled. Indeed, it's widely thought that the discovery that the analytic/synthetic distinction is unprincipled is one of the two most important achievements of modern philosophy of language; the other being precisely the theory that meaning supervenes on inferential role. If, then, we continue to assume that compositionality is nonnegotiable, it follows that one of these foundational principles of the philosophy of language is going to have to go. Which one? And with what are we to replace it?

What Now?

This is where we ought to tell you how to go about mending the cracked foundations, and we promise that we would if we could. Since we can't, however, we'll restrict ourselves to a few more or less jaundiced remarks on what we take the visible options to be and on what we think that adopting one or other of them is likely to cost.

The first possibility one might consider is to try resuscitating the analytic/synthetic distinction. There is, we think, a certain face plausibility to this suggestion. We've been seeing that compositionality and analyticity come to much the same thing so long as you accept the New Testament view that meaning supervenes on inferential role. But, in fact, it's plausible that compositionality entails analyticity *whether or not* you accept New Testament Semantics. So, for example, it's hard to see how anybody could claim that the meaning of 'brown cow' is compositional while denying that the inference from 'brown cow' to 'brown' is analytically valid. If it is undeniable that the meaning of 'brown cow' is constructed from the meanings of 'brown' and 'cow', it seems equally undeniable, and for the same reasons, that the inference from 'brown cow' to 'brown' is guaranteed by the linguistic principles that effect this construction. But an inference that is guaranteed by linguistic principles *just is* an analytic inference. In short, the very structural relations among the constituents of a sentence that ground its compositionality would appear to engender the analyticity of some of the inferences in which its constituents are involved.

So, to repeat, it appears that compositionality underwrites certain analyticities all by itself, without further appeal to the principle that meaning supervenes on inferential roles. A New Testament Semantist might thus reasonably argue that if compositionality is non-negotiable, so too is analyticity; hence that the rejection of Quine's anti-analyticity arguments is independently motivated. (It's worth mentioning in this respect that Quine says little or nothing about examples like 'brown cow' $\rightarrow$ 'brown', where the analyticity of an inference depends on the *structure* of its premises rather than their lexical contents.) But if the revival of the analytic/synthetic distinction can be defended, it looks as though everything is all right. We've

seen that you can't have New Testament Semantics and compositionality *and the rejection of the analytic/synthetic distinction*, but you certainly can have the first two if you drop the third.

We emphasize, however, that the kind of analytic/synthetic distinction that compositionality underwrites holds only between expressions and their *syntactic constituents*. It thus serves to distinguish, say, 'brown cow' → 'brown' from 'brown cow' → 'dangerous'. But it does *not* underwrite a distinction between, say, 'brown cow' → 'animal', and 'brown cow' → 'dangerous'. That is, it doesn't underwrite an analytic/synthetic distinction among inferences that turn on the lexical inventory of the premises as opposed to their linguistic structure. It is, however, precisely these lexically governed inferences that make trouble for New Testament Semantics.

If the meaning of 'brown cow' derives from the meanings of 'brown' and 'cow' (as it must do if the meaning of 'brown cow' is compositional) and if the meaning of 'cow' is its inferential role (as it must be if New Testament Semantics is right), then it must somehow be arranged that 'brown cow' inherits the inference *animal* but not the inference *dangerous* from the meanings of its constituents. But that requires that we exclude from the semantic representation of 'cow' such information as, for example, 'cow' → 'kind of *x* such that the brown *x*s are dangerous' (and, analogously, we must exclude from the semantic representation of 'brown' such information as 'brown *x*' → '*x* such that the cow *x*s are dangerous'). But to say that we must treat these inferences as excluded from the semantic representations of 'cow' and 'brown' is just to say that they are *not* constitutive of the meanings of 'cow' and 'brown'. But to say *that* is to presuppose an analytic/synthetic distinction that applies to lexically governed inferences. And it is precisely an analytic/synthetic distinction for lexically governed inferences that the familiar Quinean polemics are widely supposed to jeopardize.

We've been putting a lot of weight on the difference between analyticities that are engendered by the compositional structure of an expression and those that are generated by the meanings of items in its vocabulary. This distinction is, of course, undermined if you assume that there is a level of representation at which lexical items are semantically decomposed: to posit such a level is, in effect, to hold that lexical meaning is itself compositional. So if there is lexical

decomposition, then we can identify the meaning of an expression with its role in those inferences that are determined by its compositional structure *including the compositional structure of its lexical items*. The effect is to assimilate inferences like 'brown cow' → 'animal' to inferences like 'brown cow' → 'brown' since, at the level of semantic representation, both involve relations between expressions and their constituents. And both are distinguished from 'brown cow' → 'dangerous' since, presumably, 'dangerous' isn't a constituent of the representations of 'brown cow' at *any* linguistic level.

But, of course, to resolve the present worries by taking lexical decomposition for granted simply begs the question against Quine. If lexical meanings were compositional, then lexical items would be definable; and it's part of Quine's story about there being no principled analytic/synthetic distinction that there is no principled notion of definition. If Quine's arguments show anything, they show that there is no way to reconstruct the intuition that 'brown cow' → 'animal' is definitional and 'brown cow' → 'dangerous' isn't.

We conclude that, although there may be reasons for resuscitating the analytic/synthetic distinction, the nonnegotiability of compositionality isn't one of them. Compositionality licenses a distinction between 'brown cow' → 'brown' and 'brown cow' → 'dangerous' but not between 'brown cow' → 'dangerous' and 'brown cow' → 'animal'. So the original problem stands: if compositionality isn't negotiable, then either there is an analytic/synthetic distinction for lexically governed inferences (*contra* Quine), or inferential roles aren't what meaning supervenes on (*contra* New Testament Semantics).

Our second point is that the present situation is rife with ironies. We remarked above that once you say that the meaning of an expression supervenes on its inferential role, it's hard to stop short of saying that the meaning of an expression supervenes on its *whole* inferential role. That is, it's hard to stop short of relativizing the meaning of an expression to the whole language that contains it, with the consequences that expressions in different languages are semantically incommensurable. There is a sort of linguistic idealist who delights in these holistic implications of New Testament Semantics; you see more than a trace of this sensibility among the 'cognitive linguistics' crowd, among connectionists, and, of course, among philosophers

like Rorty, Putnam, Kuhn, and Derrida. It now appears, however, that inferential role semantics *doesn't have* any holistic implications after all, so relativism loses however the argument turns out.

Here, roughly, is how the argument from inferential role semantics to semantic holism was supposed to go:

Premise 1, New Testament Semantics. The meaning of an expression is at least partially constituted by the expression's inferential relations.

Premise 2, No Analytic/Synthetic Distinction. There is no principled distinction between those of its inferential relations that constitute the meaning of an expression and those that don't.

Conclusion, Semantic Holism. The meaning of an expression is constituted by all of its inferential relations, hence by all of its roles in a language.

There are, we think, lots of reasons to disapprove of this way of defending semantic holism; not least that it depends, apparently ineliminably, upon a form of slippery slope argument, and they are notorious for leading from true premises to false conclusions. (That is, it depends on arguing from 'there's no principled difference between the Fs that are Gs and the Fs that aren't Gs' to 'either none of the Fs are Gs or all of them are'.) For present purposes, however, we're prepared to ignore all that. The point we want to emphasize is that the argument can't be better than its second premise; that is, the argument from inferential role semantics to semantic holism explicitly depends on denying the analytic/synthetic distinction.

So, the situation is this: if the analytic/synthetic distinction is principled, then the second premise is false and the argument for semantic holism is unsound. But if the analytic/synthetic distinction is unprincipled, then the first premise is false. This is because, as we've been seeing, respecting compositionality requires that the meaning of an expression be identified with its role in analytic inferences; and compositionality isn't negotiable. But the argument that there is no analytic/synthetic distinction is the argument that there aren't any analytic inferences. We assume the law of excluded middle—either there is an analytic/synthetic distinction or there isn't. Either way, the argument from inferential role semantics to

semantic holism has to be unsound. We suppose that this demonstration should darken the skies of West Coast semantic holists. So be it; they need the rain.

The foundational problem, to recapitulate, is that you can't do all of the following: endorse inferential role semantics *and* endorse compositionality *and* abandon the analytic/synthetic distinction. If compositionality is nonnegotiable, that leaves you with only two options: endorse the analytic/synthetic distinction or reject the idea that meaning supervenes on inferential role. We can't tell you which of these it would be best to do, but we do want to insist that wriggling isn't likely to get you off the hook. In particular, lots of cognitive scientists have hoped to reconcile themselves to the Quinean arguments by opting for a graded, or a contextualized, or otherwise denatured notion of analyticity. The idea is that, although Quine may have shown that the notion of identity of meaning is in trouble, it's still open that a notion of similarity of meaning might be evolved. It's supposed that grading the analytic/synthetic distinction would allow us to have both compositionality and the supervenience of meaning on inferential role. New Testament Semantics would thus be vindicated, albeit in a sort of soft-edged version.

Nobody has shown that this can't be done; but nor does anybody have the slightest idea of how to do it. It seems to us that the arguments against a principled notion of identity of meaning work just as well against a principled notion of similarity of meaning, but we won't try to make that case here. Suffice it for present purposes to remark that the connections between analyticity and compositionality that we've been examining make the search for a graded notion of analyticity look unpromising. Compositionality is, after all, a principle that governs relations *among meanings* (it governs the relation between the meaning of a complex expression and the meanings of its constituents). So, if your semantic theory reconstructs meaning in terms of analytic inference, and if you have only a graded notion of analyticity, *then you will have to live with a graded notion of compositionality as well*. But what would a graded notion of compositionality be like? And, in particular, how would such a notion do what compositionality is required to do; viz., account for systematicity, productivity, and such? Wouldn't a graded notion of compositionality entail, at best, that a finite acquaintance with a language is ade-

quate to *sort of understand* expressions not previously encountered? Or that if a language is capable of expressing the proposition that aRb , then it is *sort of* capable of expressing the proposition that bRa ? Or that if T is the truth condition of sentence S , then the individuals and properties involved in T are *sort of* expressed by the constituents and syntax of S ? And is there any sense to be made of such claims as, for example, that:

(John's loving Mary makes 'John loves Mary' true) only if 'John' sort of refers to *John*?

These are muddy waters and we do not recommend that you wade in them.³

Here's where we've gotten to: Compositionality is nonnegotiable, and inferential roles are compositional only if the analytic/synthetic distinction is tenable for inferences that are governed by the lexical inventory. So, if Quine was right about the analytic/synthetic distinction, the only remaining alternative is to give up on the New Testament idea that the meaning of lexical items is constituted by their inferential roles. Barring some proposal for an entirely new kind of semantics, this requires going back to the Old Testament view that what makes 'dog' mean *dog* is some sort of symbol-world connection; perhaps some sort of causal or informational or nomological connection between tokens of the expression type and tokens of the animal type. There are, as remarked above, various accounts of this currently on offer and, who knows, maybe one of them can be made to work.

The spin-offs for epistemology would be of some interest since the plausible candidates for the semantically relevant symbol to world relations all look to be *atomistic*. If, for example, 'dog' means *dog* because dogs cause 'dog's, then it looks like your language could have a word that means what 'dog' means in English even if your language didn't have any other words that mean what English words do. *Prima facie*, in fact, it looks as though your language could have a word that means what 'dog' means in English even if it doesn't have any other words *at all*. All sorts of holistic arguments against the possibility of

³ Graded accounts of semantic relatedness generally take *similarity of meaning* as their fundamental explanatory construct. For discussion of what work such theories might (and might not) be able to do, see Ch. 9 of this volume.

translation (and, more generally, against the commensurability of different languages and ideologies) would then appear to be likely to collapse. More dark skies on the West Coast.

We aren't, to repeat the point one last time, actually urging you to follow this course. On the contrary, a lot of linguists will have serious reasons for not wanting to do so. For example, if Old Testament Semantics is essentially right, what becomes of such linguistic studies as 'lexical semantics'?⁴ You can't study the semantic relations among lexical items unless there *are* semantic relations among lexical items; and it's the burden of Old Testament Semantics that there aren't. Old Testament Semantics has it that semantic relations hold between lexical items *and the world* and *only* between lexical items and the world. Old Testament Semantics has, therefore, no truck with semantic fields, lexical decompositions, conceptual networks, and the like. If, indeed, what you mean by 'the semantic level of linguistic description' is the level of description at which sentences that differ only in synonymous expressions are identically represented, the natural way to read the Old Testament is as denying that there is a semantic level of linguistic description. So generative semantics goes, and interpretive semantics *also* goes, and the highest level of linguistic description is, as it might be, syntax or logical form; viz., a level where the surface inventory of nonlogico-syntactic vocabulary is preserved. The God of the Old Testament is an austere God.

As previously remarked, our present purpose isn't to sell you any such chilly theology; just to point out that there seems to be this crack in the foundations of the structure in which semanticists and philosophers of language have recently been cohabiting. And to urge that somebody do something about it not later than sometime before the roof falls in.

⁴ See Ch. 5 of this volume.

The Pet Fish and the Red Herring: Why Concepts Still Can't Be Prototypes

THERE is a standard objection to the idea that concepts might be prototypes (or exemplars, or stereotypes; we'll use these terms interchangeably): Because they are productive, concepts must be compositional.¹ Prototypes aren't compositional. So concepts can't be prototypes.²

However, two recent papers (Osherson and Smith 1988 (OS2); Kamp and Partee 1995 (KP)) reconsider this consensus. They suggest that, although the standard objection is probably right in the long run, the cases where prototypes fail to exhibit compositionality are relatively exotic and involve phenomena that any account of compositionality is likely to find hard to deal with; for example, the effects of vagueness, quantifiers, indexicals, contextual constraints, etc. KP are even prepared to indulge a guarded optimism. For example, 'when a suitably rich compositional theory . . . is developed, prototypes will

¹ A word on notation: We use italics for names of semantic values (as in 'the concept *cow* expresses the property of *being a cow*'). We use capitals for the names of concepts and for the names of 'semantic features' (as in 'the concept *cow* contains the feature +ANIMAL'). We assume that words and phrases express concepts (e.g., that the English phrase 'brown cow' expresses the concept BROWN COW), and that the productivity of language is parasitic on the productivity of thought. We are thus explicit about drawing the language/thought distinction only when it matters to the discussion. Often we'll go back and forth between concepts and terms as convenience of exposition dictates.

² For aspects of the standard objection, see Fodor (1981) and Osherson and Smith (1981). For a survey of the literature on prototype effects in categorization tasks, see Smith and Medin (1981).

be seen not as incompatible with the logic of vague concepts, but as one property among many which only when taken altogether can support a compositional theory of combination' (1995: 185).

In this paper, we argue that the standard objection to prototype theory was right after all: the problems about compositionality are insuperable in even the most trivial sorts of examples. It is therefore as near to certain as anything in cognitive science ever gets that concepts aren't prototypes; indeed, that the structure of concepts is not in any way statistical. Theories of categorization, concept acquisition, lexical meaning, and the like which assume the contrary simply won't work.

Productivity and Compositionality

We start with four assumptions that are widely granted and which we will not undertake to argue for:

1. Some of its mind-world relations are constitutive of the identity of a concept. Nothing that refers to bricks, or that fails to refer to birds, could be the concept BIRD. Practically all philosophers who discuss these sorts of issues take this for granted. (However, see Stich 1983 and Brandom 2000.) And so too do practically all psychologists, at least implicitly. For example, most of the experimental data on which the discussion of prototypes turns use categorization tasks to assess concept possession. That is, they assume that whether one has the concept C is determined, at least in part, by how well one distinguishes the things that C applies to from the things that it doesn't.

2. Some concepts are constituents of others; for example, the concepts BROWN and COW are both constituents of the complex concept BROWN COW. Likewise, concepts are the constituents of thoughts. Thus, BROWN, COW, and BROWN COW are all constituents of the thought that BESSIE IS A BROWN COW.

3. There are infinitely many concepts and (hence) infinitely many thoughts. Thus the sequence BESSIE IS A BROWN COW; BESSIE IS A VERY BROWN COW; BESSIE IS A VERY VERY BROWN COW ... is unbounded ('productive').

4. The productivity of concepts/thoughts is explained by their

compositionality: In general, the content of a complex mental representation (MR) is determined by and only by the contents of its constituent representations.³

So much for what we take to be common assumptions; we're about to see that even MR theorists who talk as though perhaps they don't accept one or another of them actually do so in practice.

We turn now to our main topic, which is how theories of the productivity of concepts fare when they take concepts to be prototypes. As remarked above, our main argument is that concepts, whatever else they are, must be compositional; that prototypes don't compose; hence that concepts aren't prototypes. It's the second premise that is crucial. We will consider (and endorse) two kinds of objections that putative accounts of the compositionality of prototypes are prone to. According to the first, prototype theory can't account for certain relations of logical equivalence among concepts. According to the second, prototype theory can't predict the semantic relations between complex concepts and their constituents.

Boolean Concepts

We start by considering the sort of complex concepts that are built up from their basic constituents by the use of such Boolean operators as AND/IF, THEN, OR, and NOT. *Prima facie*, there are two sorts of problems that such concepts raise for prototype theories. We'll argue that the first is really a red herring in that it arises *whether concepts are prototypes or not*. Correspondingly, KP's proposed solution isn't specific to prototypes: If it works at all, it does so *whatever* concepts turn out to be. But the second objection to a prototype treatment of Boolean concepts is persuasive: It appears that no such account is compatible with (4).

The Red Herring: Truth Value Gaps

Whether or not concepts are prototypes, it seems clear that many concepts are vague; that is, in many cases, no definite truth value

³ Together with its syntax. But we won't be concerned with this in the present paper.

attaches to the judgment that a certain concept applies to a certain thing. For example: it's surely just true that tables and chairs are furniture, and it's surely just false that fish are furniture; but what about wall-to-wall carpets? It's not mandatory, but it's natural, to say that there is no fact of the matter here; hence, that the judgment that wall-to-wall carpets are furniture is neither true nor false.⁴

But the fact that some (even, perhaps, all) concepts are vague doesn't, in and of itself, offer any objection to the thesis that concepts are prototypes. According to the usual formulation of prototype theory, whether something falls under a concept is a matter of *how similar* to the prototype it is. And, since SIMILARITY is itself presumably a vague concept, so too are whatever concepts are defined in terms of it. It's thus far compatible with the prototype story that judgments will often exhibit truth value gaps ('tv gaps') inherited from their conceptual constituents.⁵ So far, so good. Now what about the productivity of Boolean concepts?

If a concept exhibits a tv gap, then (assuming productivity) it should contribute that gap to infinitely many complex concepts of which it is a constituent. So, for example, if there is no fact of the matter about whether wall-to-wall carpets are furniture, then there is likewise no fact of the matter about whether wall-to-wall carpets are expensive furniture (*viz.*, whether they are both furniture and expensive). And there is also no matter of fact about whether wall-to-wall carpets are furniture or not furniture, and so forth. So, vagueness and the like make it hard to understand the productivity of Boolean concepts if it is assumed that their basic constituents are prototypes.

Whether or not you find this line of thought convincing, it bears emphasis that the problem being raised has nothing in particular to do with prototypes as such or even with vagueness as such. It comes

⁴ An alternative ('fuzzy logic') treatment assumes that there are infinitely many truth values between T and F, and that 'wall-to-wall carpets are furniture' has one of them. However, Osherson and Smith (1981) have shown that this approach leads to apparently insoluble problems in the case of concepts that are built out of the Boolean connectives.

⁵ Notice that this account of tv gaps is metaphysical, not epistemic. An epistemic treatment of vagueness would claim that there is a fact about whether wall-to-wall carpeting is furniture, but it's one that we don't happen to be appraised of; maybe future research will decide. We don't find it easy to take the epistemic view of vagueness seriously, and we won't discuss it in what follows.

up wherever you suppose that tv gaps can occur; and, *prima facie*, tv gaps can have all sorts of sources: failed presuppositions, empty names, vacuous predicates, the truth paradoxes, and so on. Likewise, we're about to see that the mechanisms that KP propose for dealing with predicates like 'wall-to-wall carpet that either is furniture or is not furniture' don't, in any way, exploit the characteristic properties of prototypes; so its success (or failure) tells us nothing about whether concepts *are* prototypes.

According to KP, the notion of *supervaluation* allows expressions with tv gaps to be constituents of thoughts that themselves *do not* have tv gaps, and so allows predicates that are vague to nonetheless be productive. We will scant the details (but see Chapter 4 for an extensive discussion of supervaluation treatments of vagueness and the like). The basic idea is that when truth conditions are assigned to complex expressions, arbitrary truth values are *stipulated* for whichever of their base clauses lack them. The truth conditions for a complex expression are then relativized to the stipulated truth values of the base clauses. Suppose, for example, that A has no truth condition because it's vague;⁶ still, the intuitively correct truth condition can be assigned not-A: it's false if A is stipulated to be true, and true if A is stipulated to be false. Notions like 'logical truths' are likewise conveniently introduced: A sentence is logically true if the proposition it expresses remains true whatever truth values are stipulated for its constituents. In such fashion, supervaluation is supposed to show how a language some of whose expressions are vague can nonetheless be compositional, thus reconciling the facts about linguistic productivity with the facts about tv gaps.

Here is the essential point, however: *The supervaluation treatment is indifferent to whether Boolean concepts are vague because their constituents are prototypes, or for some quite different reason.* For example, maybe when concepts are vague it's because mental representations are images which from time to time get smudged. Still, supervaluation continues to function, and to do so in exactly the same way as it would if concepts are vague because they are prototypes. The moral is that if vagueness is a problem, then it's everybody's problem; and if supervaluation is the problem's solution, then it too belongs to

⁶ That is to say, A has no truth value except what it may acquire by stipulation.

everybody. Which is to say: The issue about how vagueness works, and the issue about what sorts of things concepts are, are simply orthogonal.⁷

Boolean Concepts Continued

Our second point is that, whether or not prototypes generate tv gaps in Boolean concepts, there is another, and quite different, problem about assimilating the semantics of such concepts to prototype theory: For indefinitely many Boolean concepts, there *isn't* any prototype even though:

1. the primitive constituent concepts all have prototypes; and
2. the judgments in which the concept occurs all have definite truth conditions.

So, for example, consider the predicate 'isn't a cat'; and let's suppose (probably contrary to fact) that 'cat' is *not* vague; that is, 'X is a cat' has either the value T or the value F for every individual in the universe of discourse. Then, clearly, there is a definite semantic interpretation for 'X is not a cat'; it expresses the property of not being a cat, a property which all and only objects in the extension of the complement of the set of cats instantiate. However, although 'isn't a cat' and the like are entirely well behaved on these assumptions, they pretty clearly have no prototypes; and nor do indefinitely many other Boolean complex concepts. There isn't, for example, any stereotypic nonprime number, and there isn't anything that is stereotypically pink if it's square. And so on. This is a point that KP recognize explicitly (1995: 178). We remark, in passing, that this difficulty does not depend on a proprietary reading of 'prototype'; for example, it holds whether you think that prototypes are something like feature sets, as OS2 explicitly do, or whether you think that prototypes are something like exemplars. KP aren't entirely explicit about which of these

⁷ Having stressed the neutrality of the supervaluation treatment of vagueness in respect of the nature of concepts, we can't resist adding that we have pretty severe doubts whether the supervaluation treatment of vagueness actually works; in particular, whether assuming supervaluation as a semantics for tv gaps will lead to the desired result that 'John is bald or John is not bald' and the like will get appropriate truth values when their constituents have tv gaps. The problems are technical and a full treatment would go beyond the scope of the present discussion. See Ch. 4.

notions of prototype they have in mind, but on balance it seems to us to be the second.

To return to the main theme: there are indefinitely many cases in which there is no prototype corresponding to a complex Boolean concept; *a fortiori*, the MR that does correspond to such a concept isn't a prototype. Faced with this problem, a theorist might just give up and admit that, in at least indefinitely many cases, what a primitive concept transmits to its complex concept hosts is not its prototype; and that, in such cases, the identification of MRs with prototypes is simply false. The obvious elaboration of that view would be that the semantic interpretation of complex concepts is computed in the standard, classical way from their constituent structure. On that account, the predicate 'isn't a cat' has the complement of the set of cats as its semantic value because it has the conceptual structure (NOT (A CAT)). But this is, of course, to abandon the project of using prototype structure to account for the productivity of complex Boolean predicates. In particular, this treatment assumes concepts have logical forms; whereas, of course, prototypes don't.

We are seriously unclear whether, or to what extent, that is the course of action that KP endorse; their text often suggests that they have a rather different treatment in mind. Suppose you were simply to give up the idea that complex MRs have prototypes. You could then claim that, although the MR corresponding to a complex concept (e.g., NOT-F) isn't itself a prototype, its semantic value is nevertheless determined by the prototypes of its constituents.

Consider [the concept] *red*. The concept (is) *not red* does not appear to have a prototype; for how might one resolve the choice among white, green, black, yellow and all the other colors that red excludes? Nevertheless, the degree of membership in the concept *not red* is a matter of prototypicality. Only, the relevant prototype is not some prototype for *not red* but the prototype for *red*, and the degree to which something is not red is a matter of how *little* rather than how much, it resembles that prototype. (Kamp and Partee 1995: 178)

(Similarly for conjunction, disjunction, and the rest (1995: 179).) According to this treatment, if the prototype for 'red' is a fire engine in 'that's red', then it continues to be the prototype of 'red' in 'that's

not red' too. Likewise, 'red' contributes the same extension in both cases; namely, it contributes the set of things that are sufficiently similar to fire engines. However, 'red' does *not* contribute its extension to 'not red' *by* contributing the prototype for 'red' to a prototype for 'not red'; to repeat, 'not red' doesn't have a prototype.

If this is the right reading of KP, then their proposed solution to the problem of how to reconcile prototype theory with the productivity of Boolean is, in effect, to abandon the field: Since the concept RED is different from the concept NOT-RED it follows that, if the prototype for RED isn't different from the prototype for NOT-RED, at least one of the concepts must not be a prototype. Or, to put the point slightly differently, KP are proposing to abandon the identification of *the content of a constituent concept with what the concept contributes to the content of its complex hosts*. (The content of RED is given by its prototype; but *what RED contributes to NOT-RED is (not its prototype but) its extension*.) We'll consider this sort of proposal in other chapters (see, in particular, the discussion of 'reverse compositionality' in Chapter 3). Suffice it, for present purposes, to remark on how uncomfortably it sits with the general motivation of prototype theory, part and parcel of which is to *deny* that prototypes are extension determiners (see Putnam 1975; Rosch 1973; etc.). But if atomic concepts are prototypes, and prototypes don't determine extensions, then *how could it be their extensions that concepts deliver to their Boolean hosts?*

We think that prototype theory just doesn't have a story about the productivity of Boolean concepts. *Punkt*.⁸ The moral seems to be: Whether or not concepts have prototypes, what the constituents of a complex expression contribute to determining the extensions of their host is either *their* extensions or whatever it is that *determines* their extensions; nothing more, nothing less. It's not clear to us whether KP intend to deny this. For all our sakes, we hope they don't.

⁸ Notice that, *pace* KP, you can't construct the extensions for Fs of complex hosts from the set of things that are *similar* to Fs. Even transitivity of implication fails since from Fs are similar to Gs and Gs are similar to Hs, it does *not* follow that Fs are similar to Hs. Reconciling the prototype theory of concepts with the productivity of Boolean functions by abandoning the theorems of Boolean logic strikes us as thoroughly pyrrhic.

Pet Fish

There is another kind of case, discussed by both OS2 and KP, in which it is apparently not possible to provide the correct interpretation of a complex predicate given just its structure and the prototypes of its primitive constituents. The problem here is not that the complex concept fails to have a prototype, as in many of the Boolean cases, but rather than an object's similarity to the prototype for a complex concept seems not to vary systematically as a function of its similarity to the prototypes of the constituent concepts. So, for example, a goldfish is a poorish example of a fish, and it's a poorish example of a pet, but it's quite a good example of a pet fish.

Now, according to prototype theory, to have a concept is to have its prototype together with a measure of the distance between the prototype and an arbitrary object in the domain of discourse. *Prima facie*, however, the distance of an arbitrary object from the prototypical pet fish *isn't* a function of its distance from the prototypical pet and its distance from the prototypical fish. In consequence, knowing that PET and FISH have the prototypes that they do doesn't permit one to predict that the prototypical pet fish is more like a goldfish than it's like a trout or a herring, on the one hand, or a dog or a cat, on the other.

Both OS2 and KP offer solutions for this problem, but it seems to us that neither is even close to satisfactory. We will review them very briefly.

As we remarked above, OS2 take prototypes to be matrices of weighted features (rather than exemplars).⁹ So, for example, the prototype for 'apple' might specify a typical shape, color, taste, size, ripeness, etc. Let's suppose that the prototypical apple is red, and consider the problem of constructing a prototype for 'purple apple'. The basic idea in OS2 is that you form a derived feature matrix that's just

⁹ Like every other cognitive scientist we've encountered who uses the notion, OS2 don't say what a feature is, or why the concept FISH has more than the one feature +Fish. (It's true, of course, that if you ask subjects to list some typical properties of fish they will hardly ever include being fish. But familiar Gricean considerations explain their not doing so. Notice that they usually don't list *existing*, or *being things* either, though both are features that prototypical fish exhibit.) Since, however, we have long since abandoned hope of figuring out what cognitive scientists might mean by 'feature', we'll just take the notion for granted in what follows.

like the one for 'apple', except that the feature PURPLE replaces the feature RED, and the weight of the new feature is increased. 'Pet fish' presumably works the same way.¹⁰

It's pretty clear, however, that this treatment is flawed. To see this, ask yourself *how much* the feature PURPLE weighs in the feature matrix for PURPLE APPLE. Clearly, it must weigh more than the feature RED does in the feature matrix for APPLE since, though there can be apples that aren't red, there can't be purple apples that aren't purple (any more than there can be red apples that aren't red or purple apples that aren't apples). In effect, purple has to weigh *infinitely much* in the feature matrix for PURPLE APPLE because 'purple apple $\rightarrow$ purple' is a logical truth. So the theory faces a dilemma; either treat the logical truths as (merely) extreme cases of statistically reliable truths, or admit that the *weights* assigned to the features in derived matrices aren't compositionally determined even if the features themselves are. Neither horn of this dilemma seems happy. What really sets the weight of the PURPLE in PURPLE APPLE isn't PURPLE APPLE's prototype but its logical form. And, of course, prototypes don't have logical forms; they're just sets of features.

Here's another way to put much the same point: It's hard to see what the OS2 apparatus is *for*—what facts it's supposed to explain that can't be explained without it. In effect, OS2 proposes a mechanism that constructs a matrix for 'purple apple' according to which 'prototypical purple apples are purple' comes out true *even though, by assumption, prototypical apples are red*. So far as we can tell, that's all it does. But, on reflection, *there's no need* for a mechanism to do that since 'prototypical purple apples are purple' is a logical truth, as is every statement of the form 'prototypical Fxs are F (including, NB, *prototypical male nurses are males; prototypical pet fish are pets*, and so on). In short, *whether or not concepts are prototypes*, there is no need to operate on the lexical content of 'apple' in order to explain why purple apples are purple; all that's needed is the assumption that the usual principles of logic are in force.¹¹

¹⁰ Though it's not entirely clear what one is supposed to do if the prototype for FISH doesn't have a parameter for degree of domesticatedness. Maybe the pet fish features are just the union of the pet features and the fish features.

¹¹ More precisely, all that's needed is logic together with a truth that follows from the meaning of 'prototype'; viz. that prototypical Fs are Fs.

That is just as well. For, consider the inference: *a is a red apple and b is a purple apple; therefore a and b are both apples*. If 'apple' means something different in 'red apple' and in 'purple apple', why is this not a fallacy of equivocation? We think that the validity of such inferences points a clear moral: the *compositionality* of complex concepts is of a piece with the *context-independence* of the contents of their constituents. Pace OS2, it's precisely the fact that combining 'purple' with 'apple' does *not* require operating on the semantics of 'apple' (or, *mutatis mutandis*, of 'purple'; see the next section) that explains why *purple apple* → *apple* and the like are valid. These considerations bear directly on the next topic of discussion.

Feature Dependence

At best, the OS2 treatment of *purple apple* works only for cases where the 'features' associated with the Ns in AN constructions are mutually independent. This is common ground for all parties to the present discussion. So, suppose that the prototype for NURSE includes the feature *female*. You can't derive the prototype for MALE NURSE just by replacing *female* with *male*; all sorts of other things have to change too. Notice that this is true even though 'male nurse' is (in KP's term) 'intersective'; that is, even though the set of male nurses is the overlap of the set of males with the set of nurses.¹² Likewise, in both respects, PET FISH.

The most striking proposal KP have on offer is a suggestion for extending the OS2 treatment of PURPLE APPLE and the like to cases where, *prima facie*, the semantic features of a lexical item are *not* independent. Very roughly, the idea is:

1. Treat MALE NURSE (PET FISH, etc.) as analogous to BIG ANT.
2. Treat BIG ANT as meaning *big for an ant*.
3. Establish the prototype for BIG ANT by 'recalibrating' BIG for the context—ANT. This means, in effect, that the prototypical big ant is big compared to other ants; i.e., it's as good an example

¹² Things go even worse for the OS2 proposal when one considers nonintersective concepts like STONE LION, DECOY DUCK, FAKE DIAMOND, POTENTIAL PROVOST, and so forth. But we propose not to harp on this since 'antifactive' adjectives are an embarrassment for every theory of compositionality we've heard of.

of something big as an ant can be. Likewise, prototypical pet fish are as good examples of fish as a pet can be, etc.

We think this is bad advice, and not for reasons of detail.

In the first place, (1) strikes us as wrongheaded. For, whereas 'big ant' doesn't entail 'big', MALE NURSE *does* entail MALE. This asymmetry strikes us as *prima facie* an argument for not assigning them parallel analyses. Maybe BIG ANT means *big for an ant*, but MALE NURSE certainly does *not* mean *male for a nurse*; still less, *as good an example of something male as a nurse can be*. What better argument could there be that 'male' means *male* in 'male nurse' than the necessity of 'male nurses are male'? (Analogously, what better argument could there be that MALE means *male* than that 'males are male' is necessary? Indeed, what other argument have we got?)

In the second place, recalibration really undermines the whole point of compositionality. The suggestion that one has to recalibrate BIG in 'big ant' (together with the suggestion that the meaning of 'big ant' is, or includes, its prototype) implies that somebody who knows what 'big' means and who knows what 'ant' means, and who grasps the syntax, still has to *go to the world* to get the meaning of 'big ant'. Whereas, the whole point of assuming compositionality is to explain how somebody who understands the vocabulary and the syntax of a complex expression is *thereby* able to understand the complex expression itself; viz., *without relying upon further information from the world*. The moral is not, of course, that 'big ant' is noncompositional; rather, it's that grasping the meaning of 'big ant' does *not* require knowing its prototype.

Consider the following suggestion (which, we think, makes the very mistake that KP's story about recalibration does). 'Jones has an average income' means *Jones has an income that is average for an income*. So, to understand 'Jones has an average income', you have to calibrate 'average' for 'incomes'. So, you can't understand 'Jones has an average income' *unless you know what income(s) are average*. But this is clearly wrong. To a first approximation, all you need to know in order to understand 'average income' is that it's the one you get by summing all the incomes and dividing by the number of income makers. (We ignore dialects in which 'average' is a general term for

central tendencies.) You can, therefore, perfectly well know what 'average income' *means* without having the slightest idea what the average income *is*. We would guess this is the usual case.

We think PURPLE APPLE works like PET FISH (they are both intersective; i.e., $AN \rightarrow N$ is valid). We think BIG ANT works like AVERAGE INCOME (neither is intersective; i.e., AN means *A for an N* rather than *A and N*). But we think that for none of these does semantic interpretation require recalibration. Since, however, in all four cases, you *would* need to recalibrate if you were to treat concepts as prototypes, we think the moral must be that prototypes aren't, and aren't part of, the meanings of complex expressions.

The bottom line, then, is PET FISH and MALE NURSE are counter-examples, not to the compositionality of meaning, but to the compositionality of prototypes. We want to emphasize that, quite aside from the technical issues, this conclusion really does seem quite plausible. The reason that you can't derive the PET FISH prototype, given the PET prototype and the FISH prototype, is simply that what kinds of fish people keep as pets is not a fact about concepts or language. The language(/concept) assures you that the prototypical pet fish is a pet and a fish. After that, you're on your own.

Conclusion: Prototypes are Fish out of Water

The prototype for PET FISH is, as it were, an idiom; a merely linguistic(/conceptual) inquiry will tell you that pet fish are fish, but no merely linguistic(/conceptual) inquiry will tell you which pet fish are prototypical. Putting it that way might, however, suggest that there is some hope for the prototype theorist after all. Why shouldn't he just admit that his story about the compositionality of concepts doesn't work for PET FISH, but argue that the reason it doesn't is precisely that PET FISH is itself an idiom. On anybody's story, idioms are expressions where the interpretation that the compositional semantics predicts is, as it were, overridden by special conventions that must simply be acquired case by case. Having a green thumb turns out not to be having a thumb that is green, compositional semantics to the contrary notwithstanding. Why, then, shouldn't the semantics say that a

paradigm pet fish ought to be a paradigm pet and a paradigm fish? If that prediction is wrong, that only shows that 'pet fish' isn't compositional.

So, then, here's the proposal: for intersectives, the prototypical ANs are the intersection of the prototypical A things with the prototypical N things *in the unmarked case*.

The idea that composition works on prototypes to deliver default values is maybe a little counterintuitive. First blush, neither 'pet fish' nor 'male nurse' seem plausible candidates for idioms. Green thumbs are neither thumbs nor green; compare pet fish, which are both pets and fish. But it might nonetheless be a hard theory to refute. That's because it is required to predict default to the compositional prototype only in cases where no 'real world' beliefs are overriding. Failures to default can thus be viewed as *prima facie* indications that such overriding has indeed occurred. Correspondingly, a clear test of the theory would require examining cases in which there is *independent* reason to suppose that only linguistic(/conceptual) knowledge is employed in understanding an expression. It is, however, notoriously difficult to construct such cases, the vagaries of the analytic/synthetic distinction being what they are.

So, we can't prove that defaulting to the compositional prototype isn't the strategy that people actually do follow. But it's easy to see that it shouldn't be because it's an irrational strategy. That it is irrational follows from two considerations, rough formulations of which go as follows:

- (1) All else equal, the more complexly modified a concept is, the less you are likely to have special knowledge about the things in its extension.

So, for example, the present authors know a little about cows *qua* cows. But we know next to nothing about brown cows *qua* brown cows (except, of course, that they are brown and cows); and we know literally nothing about brown cows owned by people whose last names start with 'W' *qua* brown cows owned by people whose last names start with 'W' (except, of course, that they are brown, and cows, and owned by people whose last names start with 'W'). The upshot is that if our strategy were to default to the compositional prototype when we have no special information to the contrary, then

the more heavily modified a concept is, the more likely we are to default to its compositional prototype.

Notice, however, that (2) is also true:

- (2) All else equal, the more heavily modified a concept is, the less likely that its prototype is predicted by the prototypes of its constituents.

So, for example, pet fish aren't good bets for satisfying the *pet* prototype; but still less so are pet fish that live in Armenia; and still less so are pet fish that live in Armenia and have recently swallowed their owners . . . And so forth. This is actually quite close to a point we made above: the 'features' that prototypical Ns exhibit are not, in the general case, independent of one another. In consequence, the more As you put on to modify the N, the more likely it is the prototypical $A^n(N)$ s (and, for that matter, the prototypical $A^{n-1}(N)$ s) will not exhibit the features that the prototypical Ns do.

If you now put (1) and (2) together, it's clear why defaulting to the compositional prototype is an irrational strategy: on the one hand, the more complexly modified a concept is, the more likely it is that (1) will require you to default to the compositional prototype. On the other hand, the more complexly modified a concept is, the more (2) makes it likely that defaulting to the compositional prototype will give you the wrong interpretation. It is irrational to employ a strategy if the more likely you are to use it, the more likely it is to fail.

Afterthought

While this paper was being written, an article appeared by Huttenlocher and Hedges (1994), which proposes a statistical model for the formation of complex prototypes, one that relies on the assumption of independence of features. 'In constructing a conjoint category . . . the listener must make some assumption about the form of the relation between values of the constituent categories. The assumption that the categories are independent is an obvious one . . . If the assumption of independence holds, the formal mechanism we have proposed is applicable' (1994: 163). However, as we've just been

seeing, the more complex a concept is, the more likely it is that the 'listener' must depend on its internal structure (rather than his background knowledge) to decode it; and the more complex the concept is, the more the assumption that the features its constituents contribute are independent is likely not to be warranted (so long as feature assignments express statistical correlates rather than necessary conditions). The wary listener will therefore either avoid the strategy that Huttenlocher and Hedges commend, or avoid decoding complex concepts by assigning them prototypes (or, better, both).

Huttenlocher and Hedges refer to data that suggest that their subjects do in fact reliably default to the independence assumption. We suspect that if this is so, it shows only that subjects who are required to make guesses in experimental environments don't care much whether their guesses are true. It would be interesting to know what happens in situations where the outcomes matter more.

Why Compositionality Won't Go Away: Reflections on Horwich's 'Deflationary' Theory

COMPOSITIONALITY is the idea that the meanings of complex expressions (or concepts) are constructed from the meanings of the less complex expressions (or concepts) that are their constituents.¹ Over the last few years, we have just about convinced ourselves that compositionality is the sovereign test for theories of lexical meaning.² So hard is this test to pass, we think, that it filters out practically all of the theories of lexical meaning that are current in either philosophy or cognitive science. Among the casualties are, for example, the theory that lexical meanings are statistical structures (like stereotypes); the theory that the meaning of a word is its use; the theory that knowing the meaning of (at least some) words requires having a recognitional capacity for (at least some) of the things that it applies to; and the theory that knowing the meaning of a word requires knowing criteria for applying it. Indeed, we think that only two theories of the lexicon survive the compositionality constraint: viz., the theory that all lexical meanings are primitive and the theory that

¹ We assume, for the present discussion, that words express concepts, and that the content of a word is the content of the concept that it expresses. So we'll move back and forth from talk of words to talk of concepts as convenience of exposition suggests. This is, to be sure, to simplify some complicated matters. But it doesn't affect any of the questions that we disagree with Horwich about.

² See Chs. 1 and 2 in this volume, and Fodor (1998b).

some lexical meanings are primitive and the rest are definitions. So compositionality does a lot of work in lexical semantics, according to our lights.

Well, so imagine our consternation and surprise when, having just about convinced ourselves of all this, we heard that Paul Horwich has on offer a 'deflationary' account of compositionality, according to which, 'the compositionality of meaning imposes *no constraint at all* on how the meaning properties of words are constituted' (1998: 154; our emphasis). Surely, we thought, that can't be right; surely compositionality must rule out at least *some* theories about what word meanings are; for example, the theory that they are rocks, or that they are sparrows or chairs; for how could the meanings of complex expressions be constructed from any of those? What, we wondered, is going on here?

We have arrived at a tentative diagnosis, which is that Horwich fails to enforce several distinctions that turn out to be crucial. For example, sometimes he puts his main conclusion in the way we quoted above: 'the compositionality of meaning imposes no constraint at all on how the meaning properties of words are constituted'. But sometimes, even on the following page, he puts it like this: 'understanding one of one's own complex expressions (non-idiomatically) is, by definition, nothing over and above understanding its parts and knowing how they are combined' (1998: 155). Now, *prima facie*, these wouldn't seem to be at all the same theses. Whereas the first purports to answer a question about the metaphysics of *meaning*, viz., 'What linguistic facts about a complex expression are the supervenience base for its meaning properties?',³ the second purports to answer a question about the metaphysics of *understanding*, viz., 'What makes it true of a speaker that he understands an expression in his language?'⁴

³ It adds to the confusion that, although philosophers sometimes use 'how are word meanings constituted?' to ask what word meanings are, they sometimes use it to ask a quite different question, viz., what is it about an expression in virtue of which it means what it does? This last question comes up in, e.g., discussions of the 'naturalization' of semantics. Possible answers include: 'it's causal relations'; 'it's something teleological'; 'it's the communicative intentions of speaker/hearers, or the tacit conventions that they adhere to'; etc. We mention this only by way of clearing the air. It's not what either we or Horwich have in mind in the present discussion.

⁴ There are philosophers who hold, as a matter of doctrine, that nothing could be

We propose, in what follows, to consider how Horwich's deflationary account of compositionality fares if the distinction between theories of meaning and theories of understanding is properly attended to. Here's how we think it all turns out:

- Horwich is right to claim that compositionality is neutral with respect to the metaphysics of understanding expressions when 'understanding' refers to a (merely) dispositional state; *but not when it refers to an occurrent state*.
- Horwich is right *strictu dictu* to claim that compositionality is neutral with respect to the character of lexical meanings, but only *strictu dictu*.
- Compositionality taken *together with other constraints that semantic theories are required to satisfy* reduces the options in the theory of lexical meaning to the bare minimum enumerated above.

The first part of the paper is about compositionality and understanding; the second part is about compositionality and lexical meaning.

Compositionality and Understanding

Here's one version of what Horwich calls 'the basic thesis' of his paper: 'once one has worked out how a certain sentence is constructed from primitive syntactic elements, and provided one knows the meanings of those elements, then, automatically and without further ado, one qualifies as understanding the sentence' (1998: 155). We will presently argue that the question whether working out its syntax and its lexical content suffices for understanding a sentence is quite independent of the question whether compositionality constrains theories of lexical meaning. But, for the moment, we proceed to consider Horwich's account of understanding in its own right.

We think there is a (perhaps slightly forced) sense in which a theory of meaning that isn't also a theory of understanding. Thus Michael Dummett: 'Any theory of meaning which was not, or did not immediately yield, a theory of understanding, would not satisfy the purpose for which, philosophically, we require a theory of meaning' (1993: 4). Perhaps Horwich accepts this view; he doesn't say.

grasping its syntax and lexicon is indeed sufficient for understanding the meaning of a sentence or other complex expression. But we also think that there's another sense in which it's pretty clearly not. For compositionality implies that, if you are given the syntax and the lexical content of an expression, you have all the information that's relevant to what it means, hence everything you need to understand it. (Maybe there are still things that you need to know about the world; e.g., what in the world the demonstratives demonstrate. But this isn't the sort of issue that either we or Horwich are concerned with.) And we think that there is a sense of 'understanding an expression' in which having all the information that's relevant to understanding it *is* understanding. This is the notion of understanding that is in play when, for example, linguists say that, *qua* English speaker, you here and now understand infinitely many English sentences, including infinitely many whose tokens you have never encountered and never will.

However, there is also a perfectly natural sense of understanding an expression in which you can *fail* to understand one when you encounter it, *even though* you know the language that it belongs to. It's in this sense of 'understanding an expression' that one may still have some figuring out to do even after having grasped the linguistic properties on which the meaning of the expression supervenes. To see that such a situation can arise, consider sentence S, about whose syntax and lexical inventory we will now tell you the complete and unvarnished truth:

(S) 'Dogs dogs dogs dog dog dogs.'

Lexicon:

'dog_N' means *dog*

'-s' means *plural*

'dog_V' means *to dog*

Syntax:

$[[\text{Dogs}_N [\text{dogs}_{N1} [\text{dogs}_{N2} \text{dog}_{V2}]_{NP}]_{NP}]_{NP} [\text{dog}_{V1} [\text{dogs}_{N3}]]_{VP}]_S$

Clearly, someone might know everything we've just specified about its lexicon and syntax and nevertheless not understand S. Horwich discusses this sort of case, but only very, very briefly. He says, 'clearly the length and complexity of the expressions whose structures we are able to discern are constrained by psychological factors' (1998: 167).

These, however, are constraints on our 'ability to understand . . . words and appreciate how they are combined; but the compositionality of meaning is not amongst those conditions'. As far as we can make out, the idea here is that it's stuff *about our psychology* that explains our problems with S, *therefore it is not* facts about S's compositional structure. But if that is the intended argument, the premise clearly doesn't warrant the conclusion. Consider the following Silly Argument:

Silly Argument. It's stuff about your muscles that explains why you can't lift this rock, *therefore it's not* stuff about what the rock weighs.

Surely, the right answer to the Silly Argument is that it's *both* stuff about your muscles *and* stuff about what the rock weighs that explains why you can't lift it. It's *because* the rock weighs what it does that you can't lift it with the muscles you've got. Well, likewise: if you continue to have trouble with S even after we've told you its syntax and lexical inventory, that's surely because there are psychological limits that make it hard to appreciate how the syntax and lexical inventory combine to determine its meaning. To put it another way, if one continues to have trouble understanding S, that's because *an inference is required* to get from a grasp of the lexical/syntactic facts on which its meaning *supervenes* to understanding what its meaning is. This is tantamount to endorsing Horwich's 'objection 8', which goes as follows:

The deflationary account fails to do justice to the intuition that we *figure out* the meanings of complex expressions on the basis of our knowledge of what their parts mean . . . [it's] not just that the *facts* about the meanings of primitives determine the *facts* about the meanings of the complexes. It's rather that our *knowledge* of the basic facts must lead by some inferential process to our *knowledge* of what the complexes mean. (1998: 171)

To which objection Horwich replies: 'this is indeed a tempting intuition; but [the thesis that understanding complex expression requires inferences] cannot be correct, and so the deflationary attitude should not be faulted for failing to respect it' (1998: 171). We'll come in a moment to *why* Horwich says this tempting intuition 'cannot be correct'. For the moment, we want to go on a bit about

just how tempting the intuition is. An analogy should help. Consider claims (1) and (2) about checking accounts:

- (1) If you know what your balance was when you started, and what you have deposited, and what you have drawn out, then you know what the balance of your account is.
- (2) If you know the things (1) enumerates, you needn't do anything more (in particular, you needn't do anything inferential) to figure out what your balance is.

We take it that (1) is approximately truistic; it follows from *what sort of thing a balance is* (or, if you prefer, it follows from what 'balance' means⁵). Our point, anyhow, is that (2) doesn't follow from (1) and, moreover, that (2) is implausible on the face of it.

Why (2) Doesn't Follow from (1)

The following schema has the form of an intentional fallacy:

- (3) That it is the case that P determines (nominally, metaphysically, or conceptually) that it is the case that Q.
- (4) Jones knows that it's the case that P.
- (5) Therefore: Jones knows that it's the case that Q.

Accordingly, the following substitution instance of the schema is invalid:

- (6) That one started with three dollars in the account, deposited two dollars, and withdrew one dollar, determines that the current balance is four dollars.
- (7) Jones knows that he started with three dollars in the account . . . etc.
- (8) Therefore: Jones knows that his current balance is four dollars.

We suppose that (9)–(11) is likewise invalid and for the same reasons.

- (9) That 'John' means *John*, 'loves' means *loves*, and 'Mary' means *Mary* (together with syntax) determines that 'John loves Mary' means *John loves Mary*.

⁵ We don't like the parenthesized way of talking; we like to keep our metaphysics clear of our semantics. But the present issues don't in any way turn on that, so we're prepared to be concessive.

- (10) Bill knows that 'John' means *John*, and that 'Mary' means *Mary*, etc.
 (11) Therefore: Bill knows that 'John loves Mary' means *John loves Mary*.

We can now see just why, though there is 'a sense in which' it's sufficient for understanding a complex expression that one grasps its syntax and the meaning of its constituents, there is also 'a sense in which' it isn't. It's usual that what happens when P is metaphysically sufficient for Q is that the inference *believes* (. . . P . . .) → *believes* (. . . Q . . .) is valid on one way of reading 'believes' but invalid on another.

Let's see where things stand. We're pretty sure that Horwich would agree with the intuition that there's a robust reading of 'know one's balance' on which the inference (6)–(8) is fallacious. But we take it that he denies the putative analogy to (9)–(11). Why? Well because, in the linguistic case,

transitions between states of understanding do not work in this way, because the beliefs involved in knowledge of meanings are *implicit* [*sic*]. . . But since [those beliefs are] implicit, [*and thus consist*] in no more than the fact that the expression means a certain thing in his idiolect, its explanation should not be expected to involve inferential processes. (1998: 171–2; our emphasis)

So the difference between the checkbook case, where we take it that Horwich accepts the 'tempting intuition' that there is figuring out going on, and the language understanding case, where he rejects it, is that whereas the beliefs germane to checkbook balancing are explicit, the beliefs germane to sentence understanding are not. This difference matters, according to Horwich, because of a certain metaphysical truth about tacit knowledge: *Qua* explicit, your current belief that you have four dollars in the bank is constituted by your being in a mental state with certain causal powers—presumably the sorts of causal powers that affect 'transitions between states of understanding' and that are manifested when you think, talk, etc., about what you have in the bank. But, *qua* implicit,⁶ your tacit belief that 'John runs' means *John runs* is constituted simply by the fact that you take

⁶ Or perhaps it's *qua* implicit knowledge of one's own idiolect (see Horwich 1998: 172 n. 14). We're not clear whether Horwich holds a deflationary view of implicit knowledge per se, or just of implicit knowledge of language.

'John' to mean *John* and 'runs' to mean *runs* (and the syntax to be what it is). So, on this account, there's a deep difference between the metaphysics of tacit belief and the metaphysics of explicit belief.

Now, to be perfectly frank, we find this all very dark. It is, in particular, quite unclear to us why implicit beliefs should be supposed to differ, in any such way, from explicit ones. For all we know, and, certainly, for all that Horwich has argued, *implicitly* believing that *S means P* and *explicitly* believing that *S means P* might both turn out to be having your brain in a certain functional or neurological state; or they might both turn out to be having a Mentalese sentence that means that *S means P* tokened in your belief box . . . ; or whatever. If any such story is right, then it's not clear why the two kinds of beliefs mightn't be acquired by much the same kinds of inferential processes.

We can now say more clearly what our argument with Horwich is about. We think he is right that there is a kind of case (quite different from checkbook balancing) in which all that's required for grasping something complex is having the right beliefs about its structure and constituents. We take it that Horwich agrees there is a kind of case (of which checkbook balancing is an example) where having the right beliefs about its structure and constituents is *not* sufficient for grasping something complex. However, Horwich thinks the difference between the two kinds of cases is that, in the first but not the second, the beliefs about the constituents of the expression are *implicit*. By contrast, we think the difference is that, in the second but not the first, the understanding of the complex is (merely) *dispositional* (where the contradictory of (merely) *dispositional* is something like *occurrent*).

Notice that, though both apply (*inter alia*) to mental states, implicit/explicit is a quite different kind of distinction from occurrent/dispositional. The former is epistemological; it's a matter of whether the creature that's in a state has (non-inferential) access to its being there. In the simplest examples, implicit/explicit is about whether a creature is able to report being in the state that it's in. Whereas the second distinction is *ontological*; we suppose that *occurrents*, but not *dispositions*, are species of events. That is part and parcel of the fact that they are associated not just with *stretches* of time, but also with *instants*. If John's thought that the cat is trapped

in the closet is occurrent, then 'when did it occur to him?' presumably has an answer ('at 3.17', 'when he first heard the cat say meow', and so forth). But if his thought that the cat is trapped in the closet is merely dispositional (as in the case where John is congenitally disposed to occurrent *cat-in-the-closet* thoughts) the pertinent question is not 'when did it occur to him?' but 'how long did it last?' ('How lonk haf you been vorryink about die gats' beink in de gloset, Mister Portnoy?') We're aware, of course, that it's in dispute just how the distinction between dispositions and occurrents should be drawn, and we don't want to get involved in the argument. Suffice it to say that it's metaphysical rather than epistemic by general consensus; hence, by general consensus, different from implicit/explicit.

So, then, our story is that one's understanding of a sentence can be any combination of explicit/implicit with occurrent/dispositional (except that an explicit mental state presumably has to be occurrent⁷). In the case where one's understanding of an expression is implicit and (merely) dispositional, Horwich may well be right that it comes to no more than one's grasp of the syntax and the meanings of the parts. However, we think that's because such cases are *dispositional*, not because they're *implicit*. That is, like all our cognitive scientist friends, we think there is such a thing as understanding that is implicit but *occurrent* (a species of unconscious mental process, we suppose). Certainly Horwich hasn't given any reason to doubt that there's such a thing. Nor has he given any reason to believe that, when implicitly understanding a sentence is an occurrent process, inferring the sentence's compositional structure from its lexical and syntactic structure is other than essential.

We're pretty sure we have this stick by the right end since we can't think of any reason why an occurrent belief shouldn't be arrived at inferentially, whether or not it's explicit. By contrast a (merely) dispositional belief can't be arrived at *inferentially* because it can't be *arrived at* at all. There might, of course, be mental processes that cause you to have a (merely) dispositional belief, and some of the episodes in such mental processes might consist of having thoughts occur to you. But these thoughts wouldn't count as premises from which the dispositional belief is *inferred*. You can't *infer* that P unless

⁷ See n. 6.

it (implicitly or explicitly) *occurs to you* that P; and (merely) dispositional beliefs are *ipso facto* not ones that (implicitly or explicitly) occur to you.

Here's where we've gotten to now. It's urged against the deflationary account of compositionality that it 'fails to do justice to the intuition that we *figure out* the meanings of complex expressions on the basis of our *knowledge* of what their parts mean'. Not so, Horwich replies; though it's tempting to think that understanding complex expressions depends on inferences from their syntactic and lexical constituency, that thought must be resisted when the beliefs involved are implicit. Horwich offers no argument for this, however, and he needs one badly. For, the obvious candidates for non-inferential beliefs about the meanings of complex expressions aren't the *implicit* ones, they're the (merely) *dispositional* ones. If that's right, then there is no reason so far why some of one's beliefs about expressions shouldn't be both implicit and occurrent. And there is no reason so far why these *implicit*, occurrent beliefs about expressions shouldn't be much like one's *explicit*, occurrent beliefs about one's balance; viz., both inferential. We suspect, in short, that Horwich has confused the distinction between implicit and explicit beliefs with the distinction between (merely) dispositional and occurrent beliefs, and that this has led him to argue, in effect, that since when beliefs about expressions are dispositional they aren't inferential, it follows that when they are implicit they also aren't inferential. But this argument turns on a conflation.⁸

However, we're still not out of the woods. For, according to Horwich, there is an independent argument against the 'tempting intuition' that sentence understanding, like checkbook balancing, can involve a lot of 'figuring out', i.e., an argument that *doesn't* assume that implicit knowledge is *ipso facto* noninferential, but that shows all the same that 'at the fundamental level, compositionality is not explained in terms of inference' (1998: 174). Here's the argument:

although there may be *some* language whose complex expressions are understood as a product of explicit inference, such inferences would have to take

⁸ Notice that whereas (according to us) implicit beliefs may nonetheless be occurrent, it's presumably a truism that explicit beliefs can't be merely dispositional. Perhaps it's because the two distinctions fail, in this respect, to be independent that Horwich got confused between them.

place in a more basic language whose complexes would themselves already have to have some content or meaning; and if inferences are required for this, then a yet more basic language would be needed in which to conduct them . . . and so on. (1998: 172)⁹

This regress argument has been around for a long time; probably it was invented by someone who lived in a cave. The usual reply strikes us as convincing: by assumption, understanding English expressions is inferential because you have to translate English sentences into some other language (into Mentalese as it might be) in order to use the information they convey. But you don't have to translate Mentalese into some other language in order to use the information its expressions convey. All you have to do is think in it. So there isn't a regress after all.

But Horwich isn't having that.

It is all very well to refuse to speak of 'understanding' and 'possession' of meaning in connection with the language of thought, and thereby to hope to retain the idea that when a complex is, properly speaking, 'understood', inference is invariably involved . . . [But that] merely obscures the fact that the same issues [about compositionality] arise with respect to Mentalese, but in a slightly different formulation. (1998: 173)

Let's, please, be very careful about what's at issue here. To begin with, we're quite prepared to concede, for purposes of the argument, that someone who thinks in Mentalese thereby counts as understanding it 'in some sense'. *Pace* Horwich, what we've called 'the usual reply' to the regress argument is not supposed to be terminological. In particular, it's not supposed to turn on whether what one does with Mentalese counts, *strictu dictu*, as understanding it. Rather, the issue is whether our assuming that English and Mentalese are both compositional, and that understanding English is inferential, requires us to assume that whatever constitutes understanding Mentalese is inferential too. Horwich apparently thinks we must; we think we needn't, and we affirm that we don't.

⁹ There's something puzzling about this formulation. Presumably, it's a mistake for Horwich to suggest that the inferences that generate the regress have to be *explicit*. If a regress threatens, it's because of the putative necessary connection between understanding a language and making *any inferences at all*, explicit or otherwise. We'll take this reading for granted in what follows.

The following view seems to us coherent and neither vacuous nor gratuitous: English is compositional, and the process of understanding its sentences is inferential. But the fact that the process of understanding English sentences is inferential *doesn't follow from the fact that English is compositional*. So far, then, there's nothing to prevent a language from being compositional even though the process of understanding its sentences *isn't* inferential. If this possibility is coherent and begs no questions, then assuming that understanding English is inferential, *and* that all inference requires a linguistic vehicle, *and* that one thinks in sentences of Mentalese, implies no regress.

Can one conjure up a plausible story according to which all of that is true? Sure; in fact, it's just the standard language of thought story. Readers who have already heard this story, and are tired of it, are advised to skip directly to the next section.

We need, in particular, two assumptions:

- (12) The mental processes that are the consequences of understanding a sentence are mediated by a mental representation that displays its logical form; this is true both of English and of Mentalese.
- (13) Sentences of Mentalese are explicit about their logical form (in ways which, notoriously, those of English are not).

If (12) and (13) are both true, then inference might enter into one's use of English in a way that it doesn't enter into one's use of Mentalese, even though both Mentalese and English are compositional. That's because, in the case of Mentalese but not in the case of English, the truth of (13) assures that (12) is satisfied vacuously. The trick, to repeat, is to motivate the claim that understanding English is inferential *independently* of motivating the claim that it is compositional, thus leaving open the possibility that Mentalese might be compositional even though inference isn't required to use it. As far as we can tell, Horwich offers no argument at all against this tactic; it appears, indeed, that he is unaware of its existence as a polemical possibility.

Notice that we don't have to show that there is such a language in order to undermine Horwich's regress argument; all that's required is that (12) and (13) be coherent. Still, as a matter of fact, our main

reason for thinking they are coherent is that we think they are probably both true. So we'll say a little about that.

Why You Might Want to Endorse What (12) Says about English

- Because you think that understanding an English sentence involves representing it in a way that formally determines its entailments, and that logical form does so but surface structure doesn't.
- Because you think that understanding an English sentence requires (*inter alia*) recovering an ambiguity-free representation of the sentence; *a fortiori*, it requires recovering a representation which distinguishes ambiguities of logical syntax.
- Because you hold a computational view of mental processes according to which the consequences of understanding an English sentence are determined by mental operations that are sensitive to the logical form of the sentence.

There are other reasons too. But these should do to be getting on with.

Why You Might Want to Endorse What (12) Says about Mentalese

- Because you hold a view of mental processes according to which the psychological consequences of tokening a Mentalese sentence are causally determined by operations that (*a*) are responsive to its logical form and (*b*) apply to the sentence in virtue of its syntax. Making the logical form of a Mentalese sentence explicit in its syntax is how to meet both these conditions at once. As, indeed, Turing taught us.

Why You Might Want to Endorse (13)

Precisely in order to avoid the regress that threatens if you have to *infer* the logical syntax of a Mentalese sentence in order for it to play its characteristic causal role. Roughly, if (13) is true, then all that is required for it to play this role is that it be tokened (e.g., in the belief

box). That's *why* talk of understanding a Mentalese sentence is otiose in a way that talk of understanding an English sentence is not.

So then: We hold that no regress threatens the view that English and Mentalese are both compositional, and that understanding English requires a kind of inference and that thinking in Mentalese does not.

Suppose, however, that we're wrong about all this and there is, after all, some way to show that language understanding isn't inferential. What would that imply about the deflation of compositionality? In particular, what would it imply about whether the compositionality of a language constrains the semantics of its lexicon? We think the right answer is 'Nothing'.

The patient reader will remember that Horwich has two formulations of his 'main thesis', these being that 'compositionality of meaning imposes no constraint at all on how the meaning properties of words are constituted' and that 'understanding a complex expression (nonidiomatically) is, by definition, nothing over and above understanding its parts and knowing how they are combined'. We now return to a point that we made at the outset: Though Horwich asserts them interchangeably, these two formulations don't appear to be equivalent; in particular, the second doesn't appear to imply the first. Correspondingly, if it's the first that one really cares about, it doesn't matter whether the second is true. So then, after all this ground clearing, we propose to put the issues about how(/whether) understanding sentences requires making inferences entirely to one side. That allows us to turn directly to the question: 'does an account of the compositionality of a language constrain the nature of its lexicon?' We'll now argue that *of course it does*.

Compositionality and the Lexicon

We started this paper by rehearsing a familiar informal construal of the notion of compositionality: in effect, that the meanings of complex expressions supervene on their syntax together with the meanings of the lexical primitives they contain. We remarked that if a language is compositional in that sense, then, *prima facie*, that imposes some quite significant constraints on what the meanings of

its primitives could be. It's hard to see, for example, how primitive meanings could be birds or chairs since, whatever complex meanings may be, it's hard to see how birds or chairs could be parts of them.

You might suppose that Horwich would find this line of thought anathema; but he needn't, and as far as we can tell he doesn't. For, our way of putting the point assumes that the meanings of complex expressions *contain* the meanings of primitive expressions (so that the meaning of 'loves' is part of the meaning of 'John loves Mary' and so on). Such assumptions are not, however, entailed by the supervenience thesis *per se*. Supervenience *per se* entails only that *whatever the semantic facts about complex expressions may turn out to be*, they are determined by their syntax together with the semantic facts about their lexical primitives, whatever *they* may turn out to be.

Horwich's deflationary story about compositionality is only this: compositionality places no constraint on primitive meanings *if one prescind* from all assumptions about *complex* meanings except that they supervene on syntax and lexical inventory. For, as Horwich says,

whatever their underlying nature may turn out to be, there are bound to be construction properties (of the form 'x [= the meaning of the complex] results from applying procedure P to primitives whose meanings are, $\langle W_1, \dots, W_n \rangle$ '). Hence . . . it is bound to be the case that the facts regarding the meanings of the complex expressions are derived from facts about the meanings of the primitives. (1998: 160; our emphasis)

We take this to be just the point that, *if it's left open* what the semantic facts about the complexes are, then, whatever the meanings of the primitives are, there is sure to be some way of mapping the latter onto the former. If *that's* all that semantic compositionality requires, then, as Horwich says, it can't but be true.

But all that shows is that the unexiguous notion of compositionality that follows from supervenience alone isn't the robust notion of compositionality that a theory of sentence meaning requires. And, if that's all that the deflation of compositionality amounts to, it's of no great interest that compositionality deflates. People who think that compositionality substantively constrains lexical meaning have it in mind that there are all sorts of presumptive truths about the semantics of complex expressions that need

explaining; and that it's precisely the assumption of compositionality, together with a theory of the primitive meanings, that is supposed to explain them. Such presumptive truths about complex meanings as these, for example:

- Complex meanings are semantically evaluable (e.g., for truth or satisfaction).
- Although the syntax and lexicon of English are finitely specifiable, there is a denumerable infinity of distinct complex meanings.
- There are n -complex meanings for each intuitively n -ways ambiguous complex English expression.
- One meaning of the sentence 'John ate his peas' is such that 'John' has scope over 'his'.
- The meaning of 'John snores' and the meaning of 'John swims' are such that both sentences make reference to John.

And so on and on. And on.

Consider, for a further example, the arguments about systematicity that are currently live in cognitive science. Roughly, systematicity is the fact that any language(/mind) that can express(/entertain) the proposition P will also be able to express(/entertain) many propositions that are semantically close to P : Anyone who can think the thoughts that *John snores* and that *flounders swim* can likewise think the thoughts that *flounders snore* and that *John swims*. (Likewise, *mutatis mutandis* for understanding sentences of a language that can express these thoughts.)

It's pretty widely agreed that an explanation of the fact that complex meanings are systematic requires assuming that lexical meanings are context-independent. The idea is this: compositionality says that the meanings of 'John snores' and of 'John swims' depend, *inter alia*, on the meaning of 'John'. And it's because 'John' means the same in the context '... snores' as it does in the context '... swims' that, if you know what 'John' means in one context, you thereby know what it means in the other.

So compositionality, together with the systematicity of complex meanings, places a context-independence constraint on the properties of lexical meanings. This constraint is *highly substantive*. For example, it rules out the theory, held practically universally in the

cognitive psychology community, that concepts are stereotypes. The argument goes like this: the systematicity of complex meanings requires the context-independence of lexical meanings; stereotypes aren't context-independent (for example, the stereotype of people-swimming is much different from the stereotype of flounder-swimming since the latter, but not the former, adverts to the exercise of fins); so lexical meanings can't be stereotypes. See Chapter 2 for details.

This seems a pretty good example of how compositionality, together with other considerations about complex expressions, constrains the semantics of primitive expressions. Horwich, considering this case, replies that the argument from compositionality to concepts, lexical meanings, etc. not being stereotypes presupposes a 'uniformity thesis'; viz., that if the meanings of the primitives are stereotypes (or uses, or prototypes, or inferential roles, or whatever), then the meanings of the complexes are *also* stereotypes (uses, prototypes, inferential roles, etc.). Well, it doesn't, since, as we've just seen, a context-independence thesis would do equally well; and context-independence is a property that compositionality imposes on the lexicon *whether or not* uniformity is assumed.

Anyhow, there is an independent argument for the uniformity principle. Compositionality says, roughly, that its syntax and its lexical constituents determine the meaning of a complex expression; it's thus part of the explanation of why practically everybody who understands 'dogs' and 'bark' understands 'dogs bark'. But it also needs explaining that you practically never find people who understand 'dogs bark' but don't understand 'dogs' or 'bark'. What we'll call *reverse* compositionality explains this by assuming that each constituent expression contributes *the whole of* its meaning to its complex hosts. If that's right, then if you understand 'dogs bark', it follows that you know everything you need to determine the meanings of 'dog' and 'bark': in effect, the meanings of the parts of a complex expression supervene on the meaning of that expression.

Now, as far as anybody knows, compositionality and reverse compositionality *always go together*. Just as you won't find a language that can talk about dogs and barking but can't talk about dogs barking, so you won't find a language that can talk about dogs barking but can't say anything else about barking or about dogs. It would be nice

to have an explanation of why the meanings of complex expressions supervene on the meanings of their parts; and of why the meanings of parts supervene on the meanings of their complex hosts. And it would be still nicer if the explanation of these two superveniences also explained why they always turn up together. In fact, the explanation is obvious; the meaning of 'dogs bark' supervenes on the meanings of 'dogs' and 'bark' because the meanings of 'dogs' and 'bark' are parts of the meaning of 'dogs bark'; and *the meanings of 'dogs' and 'bark' supervene on the meaning of 'dogs bark' for exactly the same reason*. But the idea that the complex meanings (don't just supervene on, but) actually contain the constituent meanings is the 'uniformity thesis' in a very strong form. So it looks like the uniformity thesis must be true. So it looks like compositionality (together with reverse compositionality, together with the lack of an alternative explanation) severely constrains the lexicon after all; for example, it entails that lexical meanings can't be stereotypes.

This sort of argument ramifies in interesting ways. The meanings of 'dogs' and 'bark' must be contained in the meaning of 'dogs bark' because people who understand the sentence likewise understand the words. But the meaning of 'dogs bark' must be contained in the meaning of 'dogs bark and cats purr' because people who understand the conjunctive sentence generally understand both conjuncts.¹⁰ In fact, the reverse compositionality of complex expressions relative to their *lexical* constituents is just a special case of the reverse compositionality of complex expressions with respect to their constituents *tout court*, lexical or otherwise. Since, in natural languages, every constituent expression has infinitely many hosts, this amounts to an infinite amount of reverse compositionality, all of which is, as far as anybody knows, inexplicable unless the 'uniformity condition' is assumed. (For further discussion of the implications of reverse compositionality, see Fodor 1998b, chs. 4, 5.)

The point we want to emphasize, however, is not that the reverse compositionality argument against stereotypes as lexical meanings is correct (though it is). Our point is that people who think it matters to the lexicon whether complex meanings are compositional have

¹⁰ It's not, of course, a *necessary* truth that if you understand a syntactically conjunctive sentence you understand each syntactic conjunct; the sentence might be an idiom.

it in mind to deploy arguments that include many premises about the semantics of complex expressions over and above the assumption that sentence meanings are compositional: that sentence meanings are systematic, that languages and conceptual systems are reverse-compositional, that complex meanings are uniform with lexical meanings, etc. It's quite true, as Horwich says, that compositionality doesn't matter much lacking further such assumptions. But it's also true that it doesn't matter much that compositionality doesn't matter much lacking such further assumptions. What matters is that there appears to be a plethora of truths about the semantics of complex expressions that the assumption of compositionality, together with a good theory of the lexicon, explains; and that, as far as anybody knows, can't be explained if the assumption of compositionality is left out.

Summary and Conclusion

This was the burden of the first section: The standard reasons for holding that understanding English sentences requires making inferences are all basically 'poverty of the stimulus' arguments. They depend on claiming, on one hand, that the surface structure of English sentences is generally inexplicit about semantically salient properties of logical form; and, on the other hand, that properties of the logical forms of sentences are essential determinants of the causal consequences of their tokenings. That is all quite compatible with supposing that, although Mentalese is compositional, you don't have to understand it in order to think in it.

This was the burden of the second section: The standard arguments that run from compositionality to the nature of lexical meaning turn on the need to explain such familiar properties of complex meanings as productivity, systematicity, reverse compositionality, and the like. It is therefore not surprising, and not awfully interesting, that compositionality is deflatable when one abstracts from its role in such explanations. Why, after all, should anyone *want* to abstract from the role of compositionality in explaining systematicity, productivity, and the like? Natural languages, and human minds, *are* systematic and productive, and that they are needs explaining.

Anyhow, as far as we can tell, the argument discussed in the first section that understanding English requires making inferences, and the argument discussed in the second section that compositionality constrains lexical semantics, are independent in both directions. Thus, the compositionality argument that shows that lexical meanings can't be stereotypes applies both to English and to Mentalese, even though, by assumption, using the former requires making inferences but using the latter doesn't. Conversely, there presumably could be a language whose use has to be inferential (e.g., because its formulas aren't explicit about their logical forms) but which is nonetheless not compositional. A finite language all of whose expressions are idioms would do the trick.¹¹ The long and short is: All that the claim that understanding English requires inference and the claim that compositionality constrains lexical semantics have in common is that there are convincing arguments for each of them.

¹¹ Because he holds that compositionality places no substantive constraints theories of language, Horwich is presumably required to hold that being compositional is not a property that languages have *contingently*. (See his discussion of objection 5.) It would seem to follow that the meanings of complex expressions must supervene on their syntax and lexical contents *even in a finite language*. This strikes us as a *reductio*; surely a finite language could consist only of idioms?

What Can't Be Valued Can't Be Valued, and it Can't Be Supervalued Either

THIS is a long paper with a long title, but its moral is succinct. There are supposed to be two, closely related, philosophical problems about sentences¹ with truth value gaps (tv gaps): If a sentence can't be semantically evaluated, how can it mean anything at all? And how can classical logic be preserved for a language that contains such sentences? We are neutral on whether either of these supposed problems is real. But we claim that, if either is, supervaluations will not solve it.

There are many sentences that contain empty names; there are many sentences whose presuppositions fail, and there are many sentences whose predicates are vague (by some authors' lights all predicates are, see Kamp and Partee 1995: 129–91). It seems plausible that at least some of these sentences are incapable of semantic evaluation; *a fortiori*, that they lack truth values.² Let's suppose for a working example that 'The present king of France is bald' is one of these.

¹ For simplicity, we assume that sentences are the bearers of semantic values, meanings, and the like. It is notorious that there are more plausible candidates, but nothing in our argument will turn on this.

² It is, of course, implausible that *all* sentences with vague predicates lack truth value; 'bald' is paradigmatically vague, but 'Yul Brynner was bald' is definitely true. Generally, when we speak of vague sentences, we have in mind sentences that have tv gaps in consequence of their vagueness. For convenience, we shall call such sentences *gap-vague*.

Now, it seems that such sentences cannot be meaningless since complex sentences of which they are constituents can be logically true or logically false. For example, 'Either it is the case that the present king of France is bald or it is not the case that the present king of France is bald' looks to express a logical truth; hence, *a fortiori*, to be truth-evaluable; hence, *a fortiori*, to be meaningful. Surely, the constituents of a meaningful sentence cannot themselves be meaningless?

So, the first of our two philosophical problems is the following puzzle: On the one hand, how can a sentence be meaningful if it lacks a truth value? But, on the other, how can a meaningless sentence be a constituent of a sentence that has a truth value?

Here is the second problem: Philosophers worry that once bivalence fails and tv gaps are introduced into semantics, classical logic will have to go. In particular, for the sorts of reasons we have just been considering, they worry about the law of excluded middle. Suppose P is a sentence that lacks truth value. Then, presumably, so too does $\neg P$; and so too does ' $P \vee \neg P$ '. We confess not to be losing sleep over this, but there are purists who cannot bear it.

You might suppose that a philosopher who takes these worries seriously is going to have to insist that there just are no sentences with tv gaps after all. But Bas van Fraassen and Kit Fine, among many others, now argue that you can *both* deny the law of bivalence, thereby doing justice to the intuitions about the present king of France's being bald, *and* hold onto excluded middle, thereby placating the classical-logic freaks. Moreover, it's suggested that the mechanisms that allow you to do this make clear how sentences that actually *lack* truth values can, as it were, be meaningful in virtue of their *potential* for having them, thereby solving the first problem as well as the second.

All these glad tidings rely essentially on the method of supervaluations; with which, however, we find that we have grown disenchanted. We'll argue that there is something fundamentally wrong with using supervaluation techniques either for preserving classical logic or for providing a semantics for linguistic expressions ordinarily thought to produce tv gaps. Our discussion will focus mostly on showing that supervaluations do not solve the problem of gap-vagueness. But in reply 7, we'll make the corresponding case for tv gaps that arise from presupposition failure. Right from the start,

however, we want to emphasize that the objections we are raising are *philosophical* rather than *logical*. We have no argument with supervaluations considered as a piece of formal mathematics. But our point will be that, if you are partial to the formalism because you're worried about the two philosophical problems with which we started, you're in for being badly disappointed.

Supervaluations

So, what is a supervaluation and how exactly is it supposed to preserve classical logic and provide a semantics for vague language? We shall focus largely on Fine's³ treatment (1975: 265–300). Our exposition of this view will be as brief and as untechnical as we can manage.

We begin with 'classical valuations'. Suppose we take some first-order language *L* with monadic vague predicates; syntax as usual. And let's imagine a (partial) model *M* for *L* with assignments of extensions to predicates in such a way that some objects in the domain are *definitely* in the extension of a predicate, others are *definitely not* in its extension, and for the remaining objects it is indeterminate whether they are or are not in the extension. Semantic values for sentences will be built up in familiar ways: some sentences will be definitely true, some definitely false; and others without truth value.

A classical valuation closes the tv gaps in a partly arbitrary fashion (van Fraassen 1966a: 481–95; 1968: 136–52). On the one hand, an admissible classical valuation *M** must preserve definite truth and definite falsehood. That is:

If a sentence *S* is definitely true in the partial model *M*, then *S* is true in every admissible classical valuation *M** derived from *M* and if a sentence *S* is definitely false in *M* *S* is false in every admissible classical valuation *M** derived from *M*. (van Fraassen 1966a: 486)

On the other hand, if '*a* is *F*' is neither true nor false in *M* (because *a* neither definitely is nor is not in the extension of *F* in *M*), then the

³ Several other philosophers, however, developed supervaluation-like solutions for (some of) the problems vague language induces: David Lewis (1972: 169–218); Kamp (1975: 123–55; 1981: 225–77); and Michael Dummett (1975: 301–24; 1978: 248–68).

denotation of *a* must be quasi-arbitrarily assigned either to the extension or the anti-extension of *F* (but not both) in every classical valuation M^* of *M*—this is what Fine calls a *precisification* of the extension of *F*. In effect, it guarantees bivalence in all acceptable classical valuations of *M* (thus their name) (van Fraassen 1966a).⁴

Certain further coherence requirements—what Fine calls penumbral connections—are also imposed on these quasi-arbitrary assignments of truth values to sentences in classical valuations. For example, consider two vague predicates, 'red' and 'pink'. In the umbra of 'red' are those things which are definitely red, that is, of which 'x is red' is definitely true, while in the penumbra of 'red' are those things not definitely red, but also not definitely not red, that is, of which 'x is red' is neither true nor false; similarly for the umbra and penumbra of 'pink'. Notice that the umbras of 'red' and 'pink' exclude each other; their intersection is empty. But 'red' and 'pink' share a penumbral region assuming there are things that are borderline between pink and red.

Precisification achieves bivalence; precisified predicates have only umbras. If the precisification of a predicate is to be meaning-preserving, however, various nonlogical (in effect, 'conceptual') relations of exclusion and inclusion must be respected. In the present case, since nothing can be both red and pink, a precisification is acceptable only if the new extensions of 'red' and 'pink' do not overlap. This is what Fine calls an *external penumbral* connection (1975: 276). He also speaks of internal connections (1975: 276). If, in a precisification, Al goes into the extension of 'bald' and Bill has the same number of hairs as Al or fewer, then Bill must go into that extension of 'bald', too; and so forth.

We can now say what a supervaluation is. A supervaluation on *M* combines the results of all the classical valuations M^* on *M*; in effect, it tells us what these classical valuations have in common. Thus, a sentence is super true on a base model *M* if it is true in every classical valuation M^* of *M*; and it is super false on a base model *M* if it is false in every classical valuation M^* of *M*; otherwise, it is neither super true nor super false.

⁴ 'Thus the classical valuations do not admit truth-value gaps; bivalence holds for them. They proceed on the assumption that there is no possibility of a failure of presupposition' (van Fraassen 1969: 70–1).

So, whatever is true in the original model M will be super true in any supervaluation on M , but there may be super truths not true in the original model. For example, even if P is neither definitely true nor definitely false in M , any sentence of the form ' P or not P ' will be super true in any supervaluation on M . This is because P will be assigned either true or false in every classical valuation M^* on M ; and the semantics for 'or' and 'not' is standard. In short, even if there are tv gaps in M , because classical valuations impose bivalence, excluded middle is preserved by supervaluation. Similarly, *mutatis mutandis*, for other (propositional) classical logical truths and falsehoods.

Let's now consider how this apparatus applies to the case of vagueness in particular.

Vagueness

As we have seen, vague language is supposed to create special problems for semantics. On the one hand, if there are gap-vague sentences, bivalence is threatened since gap-vague sentences are neither true nor false. And vagueness is also supposed to threaten excluded middle when P is a gap-vague constituent of a sentence of the form ' $P \vee \neg P$ '. Supervaluation appears to cope with both these challenges.

If we say a sentence is classically valid if and only if it is true in every classical valuation on every model M , and that it is super valid if and only if it is super true on all models M , then it is provable that a sentence is classically valid if and only if it is super valid. Hence, the logic for statements containing vague predicates can be classical *whether or not* some elementary statements are gap-vague. So, supervaluation 'leads to a classical logic for vague sentences' (Fine 1975: 276).⁵

The theory is a partial vindication of the classical position. For the truth conditions are, if not classical, then classical at one remove. There is but one rule linking truth to classical truth, viz. that truth is truth in each of a set of interpretations. This rule is of general application and not dependent upon the

⁵ Cf. Kamp and Partee (1995: 149): 'the supervaluation technique . . . was motivated in large part to reflect the intuition that there is difference among such [that is, $a \ \& \ a$, $a \vee \neg a$, $a \ \& \ \neg a$] cases: if a lacks a truth value, so should $a \ \& \ a$, but $a \ \& \ \neg a$ should nevertheless be counted as definitely false and $a \vee \neg a$ as definitely true.'

nature of language or interpretation. The actual work is done by the clauses for truth in a single interpretation, and those are classical. (Fine 1975: 278)

So much, then, for our exposition. Supervaluations are alleged to provide a semantics for vague language (and, *mutatis mutandis*, language containing nondenoting singular terms), in order to preserve classical logic in the face of tv gaps (and, though we shall not discuss this, to solve the sorites (and liar) paradoxes as well). It's almost too good to be true. *More* than almost, we are inclined to think.

There are, as the reader is most likely aware, all sorts of criticisms of supervaluation techniques in the literature on philosophical logic. Some doubt that it really succeeds in preserving classical logical truths;⁶ some doubt that it can provide a solution to the sorites paradoxes (as, presumably, an account of vagueness ought to do);⁷ some doubt that it adequately acknowledges that vagueness is essential to the functioning of natural languages.⁸ We'll remain (relatively) neutral about such criticisms, partly because we do not understand a lot of them, but mostly because, as far as we can tell, none of them challenges the *intelligibility* of the idea of a classical valuation of sentences that have truth value gaps. That is what we propose to do.

What is Seriously Wrong with Supervaluational Accounts of Vague Languages

Our problem with supervaluations is that they appear to flout what one might take to be a platitude, namely:

- (P) Conceptual truths must be respected by all classical models, including classical valuations. If T is a conceptual truth, then there are no acceptable models in which $\sim T$ is true.

A word about (P): When we first wrote this paper, we took for granted that nobody would be seriously opposed to it. We have learned better since. We've been told, for example, that (P) implies

⁶ For example, von Kutschera (1975).

⁷ For example, Dummett (1975: 310–11); Sanford (1976: 206); Kamp (1981); Rolf (1984: 219–50); Wright (1987: 227–90); Sainsbury (1988); Williamson (1994: 142–64).

⁸ For example, Dummett (1975); and Wright (1987, n. 29).

an unQuinean tolerance for conceptual necessities, truths that hold in virtue of meaning alone, and the like; hence that it shouldn't be granted.

Now, it is quite true that, if you are a Quinean about meaning, you shouldn't encumber lexical semantics with notions of conceptual truth. But then, if you are a Quinean about meaning, you shouldn't be doing lexical semantics *at all*. The parties to the present discussion are all people who want a formal theory of what items in the nonlogical vocabulary contribute to the meanings of sentences (including, *a fortiori*, what vague items in the nonlogical vocabulary contribute to the meanings of sentences). We are going along with this for the sake of argument. Accordingly, we use 'conceptual truths' in (P) as a dummy term for contributions that the nonlogical vocabulary makes to meaning, whatever they may turn out to be. If there aren't any conceptual truths, then the problems under discussion don't arise. If there are conceptual truths, then they determine what the topic under discussion is, so they must not be flouted on pain of equivocation.⁹ Note that honoring (P) is the very motivation that leads supervaluation theorists to postulate 'penumbral connections'. According to Fine (and Dummett and Kamp), you need penumbral connections to ensure that (for example) there are no red things in the extension of 'pink' in any classical valuation of English. Surely, this must be because all instances of the schema 'if *x* is red, then *x* is not pink' are conceptual truths.

We endorse (P) because we think that, if there are lexical meanings and conceptual truths, then it must be that a lexical item has its meaning essentially and that conceptual truths are preserved by meaning-preserving operations. We'll say a little in defense of this later in the discussion (see especially reply 3). But we admit that these are deep issues; maybe somebody can figure out some quite different way of individuating lexical items. We are therefore content if what follows convinces you of this: you can't rely on supervaluation to solve the problems of vagueness, logical truth, and so on, unless you are prepared to abandon the intuitive connection between meaning

⁹ Tappenden (1993: 551–77) suggests that it is all right if a penumbral sentence is not true in a model, just as long as it also is not *false* in that model. Since this weak reading of P is strong enough for our polemical purposes, we do not propose to argue the issue.

and conceptual necessity. So if you do decide to go the supervalueation route, you are thereby in debt for a construal of lexical meaning that doesn't require conceptual truths to be preserved whenever meaning is. We rather think most philosophers would take that to be too high a price to pay for the benefits of supervalueation. We do; but be that as it may.

We now propose to argue that there are conceptual truths about gap-vague sentences which: (1) (P) demands that classical valuations respect; and (2) if respected, preclude certain of the precisifications that Fine, Dummett, Kamp, *et alia* say are needed if a supervalueation is to recover truths of classical logic. We take this to imply that there is no way of assigning supervalueations to gap-vague sentences which respects *all* the necessary truths, both logical and *conceptual*. So supervalueations cannot be the right semantics for gap-vague sentences.

Assume, what we suppose is plausible, that there is no fact of the matter about whether a man one ninth of whose head is covered with (his) hair is bald. (If you don't like this fraction, feel free to pick a different one; all that we require is that sentences of the form 'If x has (fraction) n of his head covered with hair, then x is bald' are definitely gap-vague.) We claim that, if there is no matter of fact about whether someone one ninth of whose head is covered with (his) hair is bald, then it is *necessary* (indeed, *conceptually necessary*) that there is no matter of fact about whether someone one ninth of whose head is covered with hair is bald. So, if it is unsettled in the actual world, then it is unsettled in every world. If you doubt this is *necessary*, ask yourself what facts about the world (or about English, for that matter) would convince you that, by gum, people one ninth of whose heads are covered with hair are definitely bald after all. If you doubt the necessity is *conceptual*, remember that baldness does not have a 'hidden essence'. You couldn't discover empirically that hairless people aren't bald in the way you could (maybe) discover empirically that XYZ isn't water.

Now consider Al. If it is noncontingent that there is no fact of the matter about whether someone one ninth of whose head is covered with hair is bald, then every model in which Al has one ninth of his head covered with hair is *ipso facto* a model in which it is indeterminate whether Al is bald. That is, the sentence 'Al is bald' (hereafter *the*

Al sentence) is indeterminate in every such model. But the point of precisification is to impose bivalence; it intends to *exclude* models in which the *Al* sentence is neither definitely true nor definitely false. To do so, it requires that we consider indefinitely many models in which *Al* has hair on one ninth of his head and the *Al* sentence is true and indefinitely many models in which *Al* has hair on one ninth of his head and the *Al* sentence is false. But now we see that there can be no models of either kind: on the one hand, the connection between having hair on one ninth of one's head and being neither determinately bald nor determinately not bald is conceptual; and, on the other, (P) tells us that acceptable models must respect all the conceptual truths.¹⁰

So you can't precisify a language that contains the sentence (not, of course, the mere form of words; a form of words can mean what you will) 'Al is bald'. So you can't precisify English. So you can't supervalue the gap-vague sentences of English, and the reason you can't is precisely that they are gap-vague. Gap-vagueness is, as one might say, an enduring trait of the sentences that have it. You can't both think it away and preserve the semantics of the sentences about which you are thinking. And, since sentences have their semantics essentially, you can't both think it away and preserve the identity of the sentences about which you are thinking. So you can't precisify a vague sentence.

Replies and Replies to Replies

We think this argument undermines the very idea of precisifying vague sentences, so we could stop here. There are, however, some interesting lines of reply that we have gleaned from the literature or heard in conversation. We find them very unconvincing; saying why

¹⁰ It has been suggested to us that conceptual truths that we want precisification to honor are somehow 'metalinguistic' or 'higher-level' in a way that the ones about 'red' and 'pink' and such are not. It's unclear to us why it would matter if this were so; and, anyhow, as far as we can see, it isn't. The crucial consideration is that you can't make a man more(/less) bald without altering his hair-to-head ratio. So if there is any valuation on which the *Al* sentence is true(/false/indeterminate) and *Al*'s head-to-hair ratio is m/n , then the *Al* sentence is true(/false/indeterminate) in every valuation in which *Al*'s head is head-to-hair ratio m/n .

we do should help clarify our view of the conceptual geography. The rest of the paper is devoted to that.

Reply 1. Someone might complain that our worry has nothing to do with supervaluation theory per se; flouting conceptual truths is something that is done all the time in classical model theory.

Suppose we ask a student to evaluate the argument: *something is a round square; so, everything is a round square*. We expect the verdict 'formally invalid in predicate logic'. Are we thereby requiring our student to do what cannot be done? Of course not! But if we can reasonably demand the consideration of models that flout the conceptual truth that nothing square is round, why can't we equally reasonably demand consideration of models in which a necessarily vague sentence has a truth value?

But this goes much too fast. To be sure, the argument we have given the student to evaluate contains a premise that is conceptually false. But that it does is *not* what makes the argument invalid in predicate logic. What makes it invalid is that it wants to infer from 'Some x is F ' to 'Every x is F '. Correspondingly, when we ask the student to evaluate the argument, we are asking him to attend only to those of its features to which predicate logic is designed to be sensitive: relations among connectives and among the quantifiers of the sentences; relations of identity and differences among their predicate structures (for example, whether the premise and the conclusion have the *same* predicate—whatever that predicate may mean), and so forth. Since, by stipulation, the argument's validity turns on *only* these features, the occurrence of 'round square' in the premise and conclusion is *inessential*. The assessment would have gone through in exactly the same way if the argument had been about red circles, or old houses, or prime numbers. When we consider the formal validity of an argument containing 'round square', we don't *flout* the meaning of the expression; we just ignore it.

So then, to repeat the question with which the objection started: If predicate logic can abstract from the conceptual incoherence of supposing a square to be round, why can't supervaluation abstract from the conceptual incoherence of supposing the A1 sentence to be definitely true? Answer: The vagueness of the A1 sentence depends on its nonlogical vocabulary; so a theory of vagueness can't abstract from

the semantics of the nonlogical vocabulary any more than predicate logic can abstract from the semantics of 'and' or 'some'. If what you care about are the 'truth conditions for a vague language' (Fine 1975: 265), you can't ignore the lexical features in virtue of which language is vague. It is the semantics of the nonlogical vocabulary that a theory of vagueness is supposed to be a theory of.

Look, if it is legitimate to save excluded middle by flouting conceptual truths that arise from the nonlogical vocabulary, why not call a spade a spade and say that what makes a sentence logically true has *nothing at all* to do with the semantics of its nonlogical vocabulary; that logical truth is determined entirely by logical form. This is what van Fraassen calls the *naïve* view that 'questions of validity can be decided on the basis of syntactic form alone' (van Fraassen 1966a: 483). Well, if it is *pretty* clever and sophisticated to ignore the necessary imprecisifiability of a vague term, why wouldn't it be *really* clever and sophisticated to ignore the rest of the semantics of the nonlogical terms as well, and hold that 'Either (you are bald if you have one ninth hair cover) or not (you are bald if you have one ninth hair cover)' is true because it is of the form ' $P \vee \sim P$ '?

Reply 2. 'You say that, if it is indeterminate whether one ninth hair cover is bald, then it is necessary that it is indeterminate whether one ninth hair cover is bald. But that can't be right. Notice that it is perfectly coherent for someone to stipulate (perhaps in a certain context or for a purpose) that one ninth hair cover is bald. On your account, that could not be so since the proposition thus stipulated would be necessarily false.'

We think this shows what we are anyhow inclined to believe: *pace* Tappenden (1993), it is unwise to take stipulation as a model for precisification. In particular, you cannot assume that, if it is all right to stipulate that (P), then there must be an acceptable model in which (P) is true. The reason you can't is that, whereas it is common ground that precisification is required to honor penumbral connections, acceptable stipulations often flout them.

We think that often enough when you stipulate that *a* is *F*, what is being mandated is not really something about the semantics of 'F', but just that *a* shall be counted with the *F*s (in the context or for the purpose at hand). When that is its force, stipulation can indeed defy patent truths, including patent conceptual truths. 'Married men

separated from their wives should count as bachelors (indeed, are bachelors) for the purpose of the census'; 'Ketchup counts as a vegetable for the purposes of the Republicans' school-lunch program'; 'Coaches count as players for the purposes of the official roster'; 'Neither Jews nor Italians count as minorities for purposes of affirmative action'; and so forth.

We have no positive account to offer of what makes, or can make, a stipulation acceptable. We would be surprised if an account of any substance could be of any generality. Suffice it that stipulations need not, and often do not, preserve the semantic features on which the identity of expressions depends. That you can stipulate that one ninth covered is bald does not, therefore, imply there is a model in which that is so.

Reply 3. Here is a positive proposal from Fine about what it is for a language to endure in the face of semantic change: definite extensions (that is, extensions in the umbra) have to be preserved.

... language can retain its identity upon precisification ... The identity of language is visible, as it were, in the permanence of recorded truth ... Let the actual meaning of a simple predicate, say, be what helps determine its instances and counterinstances. Let its potential meaning consist of the possibilities for making more precise. Then the point is that the meaning of an expression is a product of both its actual and potential meaning. In understanding a language one has thereby understood how it can be made more precise. (Fine 1975: 275–7)

The idea is that supervaluation answers the question 'How can classical logic be preserved for a vague language?' because it answers the question 'How can a sentence that cannot be semantically evaluated mean anything at all?' What preserves the meaning of a sentence under precisification *just is* that precisification preserves definite truth and falsity. Construed as an answer to our sort of objection, this reply simply refuses to niggle about the conceptual probity of assigning truth values to gap-vague sentences. Rather, it proposes a perfectly general *sufficient* condition for distinguishing precisification from stipulation (precisification preserves extensions in the umbra; stipulation does not), and it defends the condition by arguing that it draws the line at an intuitively acceptable place. In effect, this reply says that flouting (P) in the precisification of gap-vague terms is

all right because the gap-vagueness of a term is not essential to its identity:

a sentence does not become indefinite upon being made more precise. This is, perhaps, partly definitional of 'making more precise'. For what distinguishes this operation from a mere change in meaning is that it preserves truth value. (Fine 1975: 275)

Here is Dummett in a similar vein:

For every vague statement, there is a certain range of acceptable ways of making it definite, that is, of associating determinate truth-conditions with it. A method of making a vague statement definite is acceptable so long as it renders the statement true in every case in which, before, it was definitely true, and false in every case in which, before, it was definitely false. (Dummett 1975: 310–11)

The trouble with this line, according to us, is that the distinction between stipulation and precisification that it proposes is not plausible after all. We would have thought that what makes a precisification (or, *mutatis mutandis*, any linguistic change) meaning-preserving is not that it conserves unaltered the (determinate) truth values of sentences that contain the predicate, but rather that it conserves the necessary truths that contain the predicate, conceptual necessities included. Here is a thought experiment that illustrates how these two criteria for meaning preservation can diverge. Suppose that, in aid of calendar reform, we decide to skip a day; so the third of next month is a Tuesday instead of a Monday. The reform is thus unconservative about the determinate extension of 'Tuesday', but the intuition (anyhow, our intuition) is that it is conservative about what 'Tuesday' means. Now, imagine a reform that institutes a six-day week, with Wednesdays following Mondays. This violates (*inter alia*) 'Weeks have seven days' and 'Wednesday is the third day after Sunday', both of which are presumably conceptually necessary. In this case, the (our) intuition is that the meanings of the terms in this part of the language have not, to use Fine's term, 'endured the change'.

If you don't like the example, so be it; for the point it is supposed to illuminate isn't really moot in the current discussion. As remarked above, Fine and others put 'penumbral connections' in place precisely to ensure that necessities like 'if red, then not pink' are preserved

under precisification. That must be because they think precisification ought to preserve something more than determinate extension; specifically, it must be because they think that precisification ought to preserve conceptual necessities. But that it is indeterminate whether someone who has one ninth hair cover is bald is conceptually necessary, and so must be preserved under any precisification that illuminates the meaning of 'bald'; which, however, it isn't if there are acceptable models in which Al has one ninth hair cover and the Al sentence is true. Such a model fails to meet the conditions for 'enduring' language, in just the same way, and to just the same extent, as a model in which 'This red thing is pink' is true. We return to a point we made in endorsing (P): It is extremely plausible that respecting determinate extensions is not, after all, a sufficient condition for meaning to be preserved under linguistic change. This should surprise no one who thinks that meaning and conceptual necessity must somehow be intrinsically related.

Reply 4. 'All right, all right, so strictly speaking, "bald" doesn't mean "bald" in a language in which "is bald" is determinately true of someone with one ninth hair cover. So a model in which the Al sentence is definitely true isn't, strictly speaking, a model of English. But, gee, it's a model of a language that's a lot *like* English. Like enough so that we can learn interesting things about what "bald" means in English by attending to the behavior of its not-quite-English counterpart.'

Reply 3 said that the necessities whose violation precisification tolerates aren't constitutive of the meanings of gap-vague terms. Reply 4 admits that they are but is prepared to live with the resulting equivocations. The literature doesn't always respect the distinction between reply 3 and reply 4, but sometimes it does. For example, Fine says:

Under the pressure of their own use, the meanings of terms will need to change. The terms, in their old sense, will not be adequate to express the new truths, pose the next questions, make the right distinctions. Now clearly it is convenient that the changes in meaning be conservative, that the true records before the change remain true after the change. (1975: 275)

Or, consider Sainsbury:

'Heap' is vague but it could be replaced by a sharp predicate . . . Indeed such replacement is the only way forward if someone relentlessly pursues the question, asked of a penumbral object, whether it is a heap or not. We have to say: There is no answer to *that* question, but we could introduce new predicates, rather like 'heap,' of which we *could* answer that question. (1988: 32)

Or, Sainsbury again,

'Newheap' is one of many new predicates that could be introduced . . . Such a predicate will be called a *sharpening* with respect to 'heap'. . . . We will say that [a sentence containing a vague predicate] is true if and only if it is true however its vague predicates are sharpened. (1988: 33)

Oh, well, one chap's light is another chap's darkness. It is unclear to us how the vagueness of English expressions is to be illuminated by investigating the homophonic expressions in a language that is *not* English and *none* of whose terms is vague. If it is clear to you, you are welcome to it. But you owe us an account of what 'newheap' and the like are required to share with their English counterparts, if not the conceptual truths about heaps.

Reply 5. 'A model is acceptable if and only if every sentence that is determinately true is true in the model; what is the big sweat?' (Our erstwhile colleague Vann McGee once said something like this to us, only more politely.) Answer: this criterion of acceptability presupposes a viable notion of *same sentence*; it has got to be the *same* sentence that is both determinately true and also true in the acceptable model. But sentences are *semantic* entities, so a necessary condition for their identity is (presumably) the preservation of the conceptual truths into which they enter. If acceptable models can flout conceptual truths, the proposed criterion is *not* satisfied by precisifications. (And if, *per contra*, sentences are *not* semantic entities, it beats us how the fact that a sentence is true in a model if and only if a *homophonic* sentence is determinately true could have anything to do with the acceptability of the model.)

Reply 6. 'You guys think that precisifying extensions would result in, for example, the sentence "Anyone with one ninth hair cover is bald" being assigned true (that is, definitely true) in some classical valuations; but that is unsubtle. What is assigned is something like true-in-a-classical-valuation; or *better*, t-in-a-classical-valuation.

Who says a sentence can't be both necessarily indeterminate and t-in-a-classical-valuation? "t-in-a-classical-valuation" is my word, so I get to decide what it entails.'

Fine! But if t-in-a-classical-valuation is completely independent of truth *tout court*, that is, of so-called definite truth, then how can the fact that 'Al is bald' is t-in-some-classical-valuations and f-in-others make the complex sentence 'Either Al is bald or Al is not bald' definitely true? We thought the claim was that the disjunction is definitely true because at least one of its disjuncts is true in every classical valuation. If so, then why on earth should having one disjunct that is t-in-a-classical-valuation in every valuation make a disjunction definitely true?

To put the point slightly differently, the present objection is that 'true' in 'true-in-a-classical-valuation' is an orthographic accident, not to be confused with 'true' in 'definitely true'. But then how is it possible to define 'definitely true' in terms of it? Having a friend who is dogmatic is quite unlike having a friend who barks, so it would be a surprise if you could construe the one state of affairs in terms of the other. Likewise, according to the present objection, being t-in-a-classical-valuation is quite unlike being true since a sentence that is necessarily indeterminate can be t-in-a-classical-valuation. But (surprise, surprise!) being t-in-every-classical-valuation entails being definitely true all the same. How are we to understand this? And if it is not to be understood but merely to be admired, in what sense has supervaluation shown us how vague language works?

It has been suggested to us that we should think of 't-in-a-classical-valuation' as being defined not explicitly but in use (following Ludwig Wittgenstein's use of 'use' rather than Bertrand Russell's, we suppose). If we meditate on how the predicate functions across an exhaustive set of classical valuations, we shall come to understand it; and the logic and truth conditions of vague language will thereby be illuminated for us. Well, maybe. But, so far, what strikes us when we meditate on the behavior of 't-in-a-classical-valuation' is that, although we are told that it does not *mean* true, and although we take the hyphens to say that it does not *contain* 'true', still, atomic sentences that are t-in-classical-valuations contribute exactly what true sentences would to the semantics of the complex sentences that contain them. Could this be an accident?

Really, of course, for a sentence to be true in a model is for it to be *true* in that model; it is for the sentence to bear the same relation to the model that a (definitely) true sentence bears to the actual world. The intuitive force of supervaluation depends on this in all sorts of ways. For example, if you break the connection between 't-in-a-model' and 'true in a model', there is nothing left to guarantee that whatever is t-in-a-model is *ipso facto truth-evaluable*, either in that model or, indeed, at all. So, what is to stop 'The number two is feeling blue' or 'Colorless green ideas sleep furiously' from getting t-evaluated along with 'The cat is on the mat', 'The present king of France is male', and 'Al is bald'?

In fact, once you break the connection between 't-in-a-model' and 'true in a model', what makes being t-in-a-model a semantic property?¹¹ And if there is *nothing* to make it a semantic property, then presumably things that have *no* semantic properties (tables and chairs, as it might be) could be t-in-model-M, too. Isn't there something pretty clearly wrong with a theory of vagueness that is compatible with tables and chairs being vague?¹² We had supposed that supervaluation was going to distinguish between things that can't be semantically evaluated and are, indeed, not meaningful (for example, tables and chairs and 'two is blue') and things that can't be semantically evaluated but are meaningful nonetheless (for example, gap-vague sentences). We had understood that was part of what the ads commend the theory for; but apparently not.

Reply 7. Outside the base model, elementary sentences are holophrastic. (This one will take a little longer.)

Supervaluations were first applied by van Fraassen (1966a,b; van Fraassen and Lambert 1967: 225–40) to provide a semantics for free logic. Subsequently, he applied the notion of supervaluations, or a generalization therefrom, to analyze the relation of presupposition and the semantics of sentences that lack truth value on account of presupposition failures. (He then applied the conceptual apparatus

¹¹ To be sure, having a truth value is a sufficient condition for being t-(/f)-in-that-model. But that is not enough to make being t-(/f)-in-the-model itself a semantic property. Being butter is a sufficient condition for being yellow. It does not follow that yellow is a dairy product.

¹² Like classical valuations of sentences, classical valuations of tables and chairs preserve definite truth and falsity. As it happens, they do so vacuously.

to such semantic paradoxes as the liar and the strengthened liar (1969).) Since the reply we are about to consider builds on van Fraassen's account of supervaluation, and since his discussion centers not on vagueness but on tv gaps that occur when sentences contain vacuous (nondenoting) singular terms, we now turn to these. Whereas the supervaluation treatment of vagueness is widely agreed to be moot, its use in respect of problems about nondenoting singular terms has become something of a received practice.

As van Fraassen conceives of classical valuations, if the name 'Pegasus' or the definite description 'the present king of France' does not denote in a partial model M , there are no constraints on what truth values may be assigned to atomic sentences containing them in a classical valuation M^* on M .¹³ That is, truth values are assigned arbitrarily to sentences like 'Pegasus has wings' or 'The present king of France is male'. Both are assigned true in some classical valuations M^* on M , false in others, and the supervaluation story about complex sentences like 'Either the present king of France is male or the present king of France is not male' goes through as before.

But now our worry would appear to arise again. Just as some predicates are necessarily vague, some singular terms necessarily fail to denote; for example, 'the man who is taller than himself' (and also 'Pegasus' if Saul Kripke's (1980) story about names is right). So it looks like van Fraassen classical valuations are in the business of assigning (nonce) truth values to sentences whose putative referring expressions *cannot refer*; hence, to sentences which cannot have truth values. How are such assignments to be understood?

(What we take to be) van Fraassen's answer to this constitutes reply 7: what truth value a sentence has in a given M^* is independent of any semantic contents of either the predicates or the singular terms it contains. In effect, sentences like 'The present king of France is male', 'Pegasus has wings', and 'The man who is taller than himself is Russian' are all treated as mere surds; that is, as having assigned truth values but no internal semantic structure. So, for example, on van Fraassen's account, it is quite consistent for both 'The man who is taller than himself is Russian' and 'Necessarily there is no man who

¹³ Exclude identity statements; and existence statements, if there is a primitive predicate for existence.

is taller than himself' to be true in a classical valuation M^* (indeed, the second must be true in *every* classical valuation M^* since it is definitely true and classical valuation preserves definite truth/falsity). Van Fraassen wants us to take seriously the idea that no semantic value is assigned to the nondenoting singular terms in his classical valuations.

Well, of course, you can say this if you like; but, if you do, we promise not to understand you. First blush, a sentence can't have a truth value if (that is, in a model in which) its syntactically singular referring expressions do not refer. That is why people think that 'The present king of France is male' lacks a truth value in the actual world. But then, if a sentence contains a singular term that fails to refer *necessarily*, it can't have a truth value in any model. *A fortiori*, it can't have an *arbitrary* truth value in any model. So, it can't be supervalued.

Possible reply: 'No, no, you miss the point when you speak of "syntactically singular referring expressions". I am saying that "The man who is taller than himself is Russian" has no internal structure in M^* , *semantic* or *syntactic*. Or, rather, since it has no internal syntactic structure in M^* , *a fortiori* it has no internal semantic structure in M^* . As far as the assignment of a truth value in M^* is concerned, the sentence is holophrastic. Its truth value doesn't depend on its parts because it has no parts for its truth value to depend on.'

We can hardly believe our ears. By what criterion is it *the very same sentence* that does contain 'the man who is taller than himself' as a constituent when it is evaluated in M but doesn't contain 'the man who is taller than himself' when it is evaluated in M^* ? And are you telling us that what syntax a sentence has depends on what model it is evaluated in? Geez! Say it isn't so.

The point we are making about reply 7 is really continuous with the one that we made about the supervaluation of gap-vague sentences; both propose to sacrifice a kind of semantic innocence that it surely would be well to retain. In the former case, if the assignment of truth values to gap-vague sentences is allowed to violate conceptual necessities, how can it be said that the meaning of a sentence is preserved under precisification? Analogously: If a sentence with a singular term gets its truth value in one way (namely, in virtue of its compositional structure) in a model where the term denotes, but in

a quite different way (namely, arbitrarily) in a model where the term does not denote, how can it be said that the meaning of the sentence is preserved from one model to the other? What becomes of the innocent assumption that, if God were to provide us with a present king of France, he would not thereby change what 'the present king of France' *means*?

Van Fraassen says that 'a classical valuation gives a faithful picture of which sentences are true and which false in some (possible) situation' (1969: 72). Faithful to what, we wonder? Not to truth values since they are assigned arbitrarily. Not to internal semantic structure, since it is ignored. Not to internal syntactic structure, since it is ignored, too. Not to truth conditions, since, *prima facie*, some of the sentences with which we are concerned ('The man who is taller than himself is Russian') are necessarily indeterminate. All that the analysis is really faithful to, as far as we can tell, is the logical form of the complex sentences in which nondenoting expressions occur. See above, reply 1 and reply 4, for the corresponding point about vagueness.

No English sentence of the form 'The man who is taller than himself is X' can be true. This is conceptually necessary, so it is true in every model in which the semantics of English singular terms is respected. To insist that there is a model in which some *other* sentence—one that *doesn't* contain the expression 'the man who is taller than himself'—is true is to tell us nothing about the semantics of singular terms in English. It tells us just that you have decided to change the subject.¹⁴

¹⁴ Like us, Ermanno Bencivenga has misgivings about van Fraassen's not making the evaluation of elementary sentences depend on internal structure. Bencivenga remarks, for example, that the identity symbol is as good a logical constant as the disjunction symbol, but the former will be invisible if the internal structure of elementary sentences is ignored.

We entirely approve of Bencivenga's attempt to avoid holophrasism, but his theory cannot avoid our chief worry, which is how to avoid valuations that assign referents to necessarily nonreferring expressions while also avoiding valuations that change the subject. A full treatment would need more space than we have, so we shall scant the details. (Although his theory has altered slightly over the years, we shall restrict discussion to the following: 1980a: 219–29; 1980b: 393–405; 1991: 98–11). Skyrms has independently advocated nonholophrastic evaluations (1968: 477–82; 1970: 153–61). We shall not discuss Skyrms, but we believe our criticisms apply to his proposal, too.)

Bencivenga's assignment of denotations to definite descriptions is perfectly

Conclusion

The intuitive appeal of the supervaluation technique comes, we think, from cases like 'The present king of France is male'. Since it is only contingent that there is currently no king of France, it is only contingent that this sentence lacks a truth value. Where a tv gap is merely contingent, we know pretty well what it would be like to fix the world so as to make the gap go away. The situation is less clear in the case

orthodox: in any model, 'the F' will have for its denotation the unique member of the domain that satisfies 'F', if there is one. If there is no such unique member in a model, then a sentence containing 'the F' will lack a truth value in that model. Since complex sentences are evaluated in the usual way, the truth value even of cases of the excluded middle, like 'The winged horse is white or the winged horse is not white', will be undefined in *some* models. This means that Bencivenga cannot hold that a sentence containing 'the F' is super true just in case it is true in every extension of the base model M. Instead, he will restrict the quantification over models just to those extensions in which a value for the sentence is defined. This spares him the need to assign arbitrary denotations to empty descriptions.

In effect, what Bencivenga does is this: as usual, let a model M be a pair $\langle D, f \rangle$, where D is the domain and f a function that assigns extensions to terms in the familiar ways. If $M = \langle D, f \rangle$ and $M^* = \langle D^*, f^* \rangle$, then M^* is an extension of M just in case D is a subset of D^* and f is a subset of f^* . Since there are no winged horses in any customary model M for English, the intersection of $f(\text{'winged'})$ and $f(\text{'horse'})$ will be empty in all the customary models, and the denotation of 'the winged horse' will accordingly be undefined in all the customary models.

It is easy to imagine, however, a model M^* that is an extension of M in which: $f^*(\text{'winged'}) = f(\text{'winged'})$, $f^*(\text{'white'}) = f(\text{'white'})$, and $f^*(\text{'horse'}) = f(\text{'horse'}) + 1$ white seagull. For such an M^* , $f^*(\text{'the winged horse'}) =$ that seagull and 'The winged horse is white' is true in the model. This model is *ipso facto* relevant to the supervaluation of 'The winged horse is white'. So too, of course, is a model in which $f(\text{'horse'})$ is extended to include a black crow rather than a white seagull and 'The winged horse is white' is false.

So there are some extensions of the customary model in which the value of 'The winged horse is white' is defined and the sentence is true and there are some extensions in which the value is defined and it is false. It follows that this sentence is neither super true nor super false in the customary model. But 'The winged horse is white or the winged horse is not white' will be super true in the customary model since it is true in every extension where the value of 'The winged horse is white' is defined. (This is, in fact, a considerable simplification of Bencivenga's treatment, but it conveys the underlying intuition. It will suffice for our purposes since his notion of extension is less restrictive than the one we have just described.)

The reader will, no doubt, have anticipated our objection: if all that is required of $f^*(\text{'horse'})$ is that it include $f(\text{'horse'})$ as a subset, then extensions will not, in general, be meaning-preserving. Patently, they will not be if $f^*(\text{'horse'})$ includes elements to which it is *conceptually necessary* that 'horse' does not apply. Consider the extension in which 'The winged horse is white' is true because $f^*(\text{'horse'})$ contains a seagull.

of sentences with vague terms, because it is less clear that it is the world's fault that they lack truth values; maybe you can't make them precise without changing just the thing about them that you are trying to study.

That certainly seems to be so when lack of truth value is conceptually necessary; presumably, conceptual necessity is itself a semantic phenomenon, at least *inter alia*. Which is to say that it must be preserved by any operation that claims to be meaning-preserving. For a semantic theory to fool around with the *constitutive* properties of an expression is for it to describe the semantics of some *other* expression. To solve a semantic problem by pretending away necessary truths is only to pretend to solve the problem.¹⁵

Appendix 4.1 Reply to Critics

Michael Morreau has written a reply to us (1999: 148–56) and, frankly, we are *completely* bewildered by his solution to the puzzle we offered. This isn't a rhetorical flourish; either we don't understand him at all, or, if we do, then his proposal isn't even of the right *form* to be a solution.

The point is entirely straightforward. The heart of Morreau's article is the

There would seem to be only two possibilities: either it is a model in which some seagulls are horses, or it is a model in which 'horse' does not mean 'horse.' Since we are inclined to think that it is some sort of necessary truth that seagulls are not horses, we think that the evaluation of 'The winged horse is white' as true in this model must involve an equivocation on 'horse'. (If, however, our modal intuitions strike you as wrong, the point can be adjusted to whatever ones suit your fancy. For example, let the sentence to be evaluated be 'The prime horse is equal to 2' and let f^* ('horse') = f ('horse') + 2.)

If 'horse' doesn't mean 'horse' in these sorts of models, then 'the winged horse' doesn't mean what it customarily means; that is, it doesn't mean what it means in the models that represent English. In that case, it is hard to see how a notion of super truth that is defined relative to such models could be germane to whether English sentences of the form 'The winged horse is F or the winged horse is not F' are true. This is, of course, the same dilemma that we kept encountering in the body of the text: where tv gaps are necessary, the choice supervaluation offers is to ignore salient structure or to equivocate. Neither alternative strikes us as awfully attractive.

¹⁵ We would like to thank Charles Chihara, William Craig, Michael Dummett, Kit Fine, David Lewis, Vann McGee, Brian McLaughlin, Terry Parsons, Mark Sainsbury, Stephen Schiffer, Michael Tye, Bas van Fraassen, and Ed Zalta for comments on earlier drafts; we especially want to thank Lou Goble, with whom we have had an extended and ongoing conversation and correspondence.

following claim: The sentence 'For each x , if x is a man and one ninth of x 's head is covered with hair, then it is indeterminate whether x is bald' is true iff 'there are two kinds of complete specifications: some in which men one ninth of whose heads are covered with hair fall within the positive extension of "bald", and others in which they fall within the negative extension'. But, we claim, this condition cannot be unvacuously satisfied; our worry was precisely that if 'Fa is indeterminate' is a conceptual truth (which Morreau is prepared to concede in the present case), then the notion of an assignment of a truth value to 'Fa'—hence of the assignment of a to either the negative or the positive extension of F —makes no sense. There is no answer to the question 'what would it be like if a necessarily indeterminate sentence were true(/false)?', just as there is no answer to the question 'what would it be like if a necessarily true sentence were indeterminate?' Necessary truths can't have even *possible* counterinstances; that's what makes them necessary. *A fortiori*, necessarily indeterminate sentences can't have precisifications that assign them truth values. Assigning a truth value to a necessarily indeterminate sentence is like assigning one to a chair.¹⁶

We think Morreau (and others) think that we think that if P has a truth value under a precisification, then it follows that P is possibly true; and they think that we haven't noticed that we think this. Well, we do think it, and we have noticed. Of course, 'precisification' is their word, and they are free to define it in a way that makes 'true under a precisification' *not* entail 'possibly true'. But we doubt that they can do so consonant with the definition being intelligible (consonant with 'true', 'possible', 'indeterminate', 'necessary', etc. meaning what they usually do). In any case, it begs the question against our objection to define 'precisification' by a construction which involves, *inter alia*, assigning truth values to necessarily indeterminate sentences. " p is true(/false)" under a precisification' doesn't, of course, entail ' p is true(/false)'. But it does, self-evidently, entail ' p is *possibly* true(/false)'. If p isn't even the kind of thing that is possibly true or false, what on earth could it mean to require (e.g.) that those of its entailments which are conceptually necessary are preserved under precisification?

Morreau (and McGee and McLaughlin unpublished) both stress that,

¹⁶ Here's a recipe for precisifying chairs (in, we think, exactly the same sense that supervaluation provides a recipe for precisifying necessarily indeterminate sentences): Assign T to every chair that is determinately true and assign F to every chair that is determinately false. If a chair is neither determinately true nor determinately false (there are, in fact, quite a lot of those), assign it T or F at random. If a certain chair is T on some assignments and F on others, then its truth value is indeterminate.

Did we hear someone object: 'Not fair; chairs aren't the *kind* of things that can be (that can be said to be) either T or F?' 'Quite so,' we reply; 'but neither are necessarily indeterminate sentences.'

according to supervaluation treatments, indeterminacy is a *global* property of sentences. For a sentence to be indeterminate is, by definition, for it to be true on some precisifications and false on others, so that a sentence is indeterminate does not entail that there is any precisification on which it *lacks* a truth value. Quite so. But that doesn't show that the notion of the precisification of a necessarily indeterminate sentence is coherent. At most it shows that the notion of the precisification of a necessarily indeterminate sentence is coherent *if the notion of a necessarily indeterminate sentence being assigned a truth value is coherent*. Which, however, is exactly what we doubted. To repeat: we can't make any sense of the claim that a necessarily indeterminate sentence has a truth value *at all*. *A fortiori*, we can't make any sense of the claim that a necessarily indeterminate sentence has a truth value according to a certain precisification. Just stipulating that it does doesn't help, of course, since it's plausible that such a stipulation would require a meaning change in one or more of the key terms and hence, eventually, in the notion of vagueness that precisification is used to define. An analogy to 'semantically anomalous' sentences may help. Is there an 'acceptable' precisification of English on which 'colorless green ideas sleep furiously' is true?

Morreau thinks that we think that supervaluation treatments won't let one assign *indeterminate* (tv gap) to any sentence; and he thinks we think this because he thinks we want to argue: 'Since every precisification assigns a truth value to each sentence, there are no tv gaps according to supervaluation treatments'. But that argument is fallacious, because just as Morreau says, indeterminacy is a property of *sets* of precisifications, not of precisifications per se. Right, except that wasn't our argument. Our point is that in the course of deciding whether a given sentence has the global property in question, we are being asked to do something that we can make no sense of; namely, to assign a truth value to a sentence that, as a matter of conceptual necessity, can't have one.

Likewise, if 'Fa' is necessarily indeterminate, we can't make any sense of the notion of assigning the individual *a* to the extension F. This is really the same point we made two paragraphs back. By our lights, if it makes sense to speak of the assignment of *a* to the extension F, then there is a possible world in which *a* is F. But if there's a possible world in which *a* is F, then it isn't conceptually necessary that 'Fa' is indeterminate. Apparently Morreau and McGee and McLaughlin think that we think that supervaluation treatments are *inconsistent*. We don't. We think that supervaluation treatments are *unintelligible*; the question of consistency doesn't so much as arise.

Part II

Decomposition

The Emptiness of the Lexicon: Reflections on Pustejovsky

A CERTAIN metaphysical thesis about meaning that we will call inferential role semantics (IRS) is accepted practically universally in linguistics, philosophy, and the cognitive sciences:¹ the meaning (or content, or 'sense') of a linguistic expression² is constituted, at least in part, by at least some of its inferential relations. This idea is hard to state precisely, both because notions like metaphysical constitution are moot and, more importantly, because different versions of IRS take different views on whether there are constituents of meaning other than inferential role, and on which of the inferences an expression occurs in are meaning-constitutive. Some of these issues will presently concern us, but for now it will do just to mention such familiar claims as that: it's part and parcel of 'dog' meaning *dog*³ that the inference from 'x is a dog' to 'x is an animal' is valid; it's part and parcel of 'boil' meaning *boil* that the inference from 'x boiled y' to 'y boiled' is valid; it's part and parcel of 'kill' meaning *kill* that the

¹ By a 'metaphysical thesis about X' we simply mean a thesis about which properties of Xs are essential. By stipulation: If theory T holds that it is metaphysically necessary that Xs are Ys, then T says that being Y is metaphysically necessary for being X, and that being Y is 'constitutive' of X-ness.

² The corresponding doctrine is generally supposed to be true of the meaning (/content/sense) of mental entities like concepts and beliefs. Our discussion is intended to apply, *mutatis mutandis*, to either version of IRS.

³ Cited forms in quotation marks, and expressions in italics either stand for canonical names of meanings or are emphasized expressions. We use capital letters for canonical names of semantic representations.

inference from 'x killed y' to 'y died' is valid; and so on. (See Cruse 1986, ch. 1 and *passim*.)

IRS brings in its train a constellation of ancillary doctrines. (Whether these ancillary doctrines are literally entailed would depend on how IRS is formulated.) Presumably, for example, if an inference is constitutive of the meaning of a word, then learning the word involves learning that the inference holds. If 'dog' means *dog* because *dog* $\rightarrow$ *animal* is valid, then knowing that *dog* $\rightarrow$ *animal* is valid is part and parcel of knowing what the word 'dog' means; similarly, knowing that *x boiled y* $\rightarrow$ *y boiled* is valid is part and parcel of knowing what 'boil' means; and so forth.

IRS also constrains grammatical theories since, on standard versions of the IRS view, the *semantic lexicon*⁴ of a language is supposed to be the component of a grammar that makes explicit whatever one has to learn/know to understand the lexical expressions of the language. IRS thus implies that meaning-constitutive inferences are part of the *semantic lexical entries* for items that have them. Lexical entries are therefore typically *complex* objects ('bundles of inferences') according to standard interpretations of IRS. It is this latter thesis that will primarily concern us in the present discussion. For reasons we have set out elsewhere, we are inclined to doubt that IRS can be sustained (Fodor and Lepore 1992). So, if IRS is the argument for the complexity of lexical entries, we are inclined to doubt that lexical entries are complex. Correspondingly, the primary question will be *whether there are any persuasive arguments for the complexity of lexical entries that do not presuppose IRS*. Our main interest in Pustejovsky (1995) is that it purports to provide such an argument.⁵

Here is how we will proceed: The natural alternative to the claim that lexical entries are typically complex is the claim that lexical entries are typically atomic (i.e., that they lack internal structure). We propose to adopt a version of this claim as a sort of null hypothesis: The only thing a lexical entry specifies is the denotation of the item it describes. Here again we scant the details for the moment. Roughly,

⁴ In what follows, 'semantic lexicon' and 'lexicon' are interchangeable unless we are explicit to the contrary.

⁵ A number of people, including an anonymous reviewer, have asked us why we have singled out Pustejovsky's theory in particular. The answer is that many of our objections to Pustejovsky's theory do not depend on the details and apply to other current popular accounts of the lexicon as well.

though: the lexical entry for 'dog' says that it refers to dogs; the lexical entry for 'boil' says that it refers to boiling; and so forth.⁶ We will try to show that all the standard arguments for rejecting this null hypothesis, Pustejovsky's included, either depend on assuming IRS or are independently unsound.

In the first two sections we consider Pustejovsky's account of the semantic lexicon; in particular, we discuss and reject his argument that the complexity of lexical entries is required to account for *lexical generativity*. In the third section we consider whether our default theory, lexical atomism, might not actually be true. We argue that, strictly speaking, it probably is not; strictly speaking, lexical *entries* are typically complex. But we claim that they are complex in a way that does not jeopardize either the thesis that lexical *meaning* is atomistic, or the identification of lexical meaning with denotation.

Pustejovsky's Theory

Pustejovsky's (hereafter 'JP's') most interesting arguments for lexical complexity don't depend upon assuming IRS, but we take the theory of the lexicon that he holds to be firmly within that tradition. In particular, like other proponents of IRS, JP thinks that word meanings are constituted by inferences, hence that knowing what a word means involves knowing (some of) the inferences in which it participates. In fact, he apparently thinks (what IRS does not strictly require) that understanding a token (e.g., an utterance or inscription) of an expression involves *actually drawing* some of the inferences that are licensed by what the expression means: 'The structuring of . . . taxonomic information . . . is not simply an exercise in domain modeling; it is necessary for driving *the inferences that a language reasoning system must perform in order to understand a sentence*' (1995: 19; our emphasis). As usual with IRS-motivated theories of meaning, it is taken for granted that the semantic lexicon should somehow formally specify

⁶ If it's assumed that it's necessary that dogs are animals, then, of course, whatever denotes a dog denotes an animal. But, according to the present view, the lexical entry for 'dog' does not provide this information; it says only that 'dogs' denotes dogs. Correspondingly, according to the present view, knowing that dogs are animals is not necessary for knowing what 'dog' means. It is, of course, not possible on any view that knowing the meaning of a word requires knowing all of the necessary truths that it can be used to express. See below.

the inferences by which meaning(/knowledge of meaning/sentence understanding) is constituted.

We stress this because JP occasionally writes as though it were not the specification of *inferential roles* but rather the specification of *denotations* with which lexical semantics is primarily concerned: 'Lexical semantics is the study of how and what the words of a language denote' (1995: 1). But this is a little misleading. Roughly, according to JP, lexical entries specify denotations via their meanings, so that, for example, coextensive expressions may well be assigned distinct lexical entries.⁷ Thus, 'John bought the book from Mary' and 'Mary sold the book to John' are made true by the same event; but they differ, according to JP, in a property of 'headedness' or 'focus' that they inherit from the lexical entries of their respective verbs. Patently, however, 'headedness' is not a property of denotations ('things in the world') but rather of denotations *as represented*. JP says that 'Headedness is a property of all event sorts . . .' (1995: 72); but he can't really mean that. What he must really mean is that it's a property of all (semantically well-formed) representations of events (in English).

Similarly, one of JP's characterizations of an 'event structure' specifies an event *e*₃ such that 'there is no other event that is part of *e*₃' (1995: 69). But, clearly, all events have parts (down to whatever physics determines is the microstructure of the universe). What JP means is really that no other event is *represented as* (or implied as) part of *e*₃ by the form of words that expresses the event. Here again, what the semantics is telling us about is not (or not just) what a word represents, but how it represents it; in effect, not just the word's denotation but also its sense.

We don't mean to insist on what are, probably, not confusions but just JP being casual about the use/mention distinction. However, it is important to us to contrast JP's sort of project with a bona fide denotational atomism, according to which satisfaction conditions, and

⁷ Whether JP thinks that the semantic representations that the lexicon specifies actually *determine* denotations (in the way that sense is traditionally supposed to determine reference) is unclear from the text, though there are passages that suggest he does. For example, discussing the 'qualia structures' that constitute the lexical entries for nouns, he says that they 'contribute to (or, in fact, determine) our ability to name an object with a certain predication' (1995: 85).

properties defined in terms of them, are the only semantic features of linguistic expressions that lexical entries specify. In fact, JP enumerates a variety of constraints on semantic theories that a purely denotational lexicon clearly could not meet, hence that are supposed to motivate a richer notion of lexical semantic representation than a denotational lexicon could provide. With one exception, these are familiar from discussions in lexical semantics dating as far back as Katz and Fodor (1963). Consonant with our overall project, we propose to discuss them relatively briefly, putting aside the ones that are motivated either by IRS or by other assumptions that we regard as tendentious. This will leave for the next major section ('Generativity') to consider what we take to be JP's main line of argument and the main contribution of his book: the claim that only if lexical entries are typically complex can the generativity of the lexicon be captured; *a fortiori*, that an atomistic lexicon would fail to explain 'how words can take on an infinite number of meanings in novel contexts' (1995: 42).

As far as we can make out, other than the considerations about generativity, JP offers three kinds of arguments for complex lexical representations; we take these up in the next three subsections.

Interlexical Semantic Relations

Lexical semantics is required to specify 'how words are semantically related to one another' (Pustejovsky 1995: 23), including, in particular, relations of 'synonymy, antonymy, hyponymy and lexical inheritance, meronymy, entailment and presupposition'.⁸

Clearly, a lexical entry that says only that 'dogs' refers to *dogs* will not thereby specify (e.g.) that dogs are animals (hyponymy); a lexicon that says only that 'bachelor' refers to *bachelors* will not thereby specify that 'bachelor' means the same as 'unmarried man' (synonymy); and so on. In fact, barring appeal to meaning postulates (of which JP disapproves; see 1995: 54, 110), the only way of capturing such relations in the lexicon would seem to require complex lexical entries (e.g., ones that include ANIMAL in the entry for 'dog')

⁸ JP says nothing about how to decide which lexical relations are on the list of interlexical relations to which grammars must be responsive, or whether the enumeration he provides is supposed to be complete.

and UNMARRIED and MAN in the entry for 'bachelor'). In short, if the lexicon is constrained to capture 'interlexical relations', then it is neither atomistic nor purely denotational.

The question, however, is how JP is to motivate imposing this constraint on lexical entries, and here the problems are formidable. The basic issue is this: all the relations JP enumerates presuppose a notion of *analyticity* (i.e., lexical entailment); so, in effect, JP requires that the semantic lexicon reconstruct the notion *analytic inference*. But this notion is notoriously problematic, and not just for the familiar Quinean reasons. Arguably, indeed, there simply is no such notion; if so, then the program of lexical semantics that JP advocates cannot be carried out. This is, to be sure, a large issue, and it has implications for many projects other than JP's. Suffice it, for our present purposes, to stress just two points.

First, Quine's own doubts about analyticity are part of his larger skepticism about necessity. But the two need to be distinguished since analyticity does not follow even if necessity is granted. Suppose, for example, that *dog* $\rightarrow$ *animal* is necessary; that is, suppose nothing could be a dog that wasn't an animal. Remember that semantic lexicons are supposed to be repositories of (and only of) meaning-constitutive inferences; that is, of those inferences that one has to be prepared to draw on pain of not understanding the word in question. But it seems clear, given this condition, that the *necessity* of an inference is not, per se, sufficient for its meaning-constitutivity.⁹ *Two is prime* is necessary if anything is, but it's implausible that you can't know what 'two' means unless you know that two is prime. Similarly, it's presumably necessary that what is square isn't circular, but it's implausible that you can't know what 'square' means unless you also know what 'circle' means. Such examples are legion.¹⁰ The moral is that even if we had a theory of necessity, it wouldn't, all by itself, give us a theory of synonymy, antonymy, hyponymy, and the like; it wouldn't even certify the *possibility* of such a theory since, even if it were perfectly certain that the distinction necessary/contingent is

⁹ See n. 7.

¹⁰ Transitivity severely compounds the problem. Since it's necessary that what's circular isn't triangular, it follows that you can't know what either square or circle means unless you know what triangle means. And so on *ad infinitum*.

principled, that would not be a reason for believing in analyticity (or, *a fortiori*, in synonymy, antonymy, hyponymy, and the like).

Second, and worse, even if analyticity is real, it is perfectly possible to doubt that it is connected, in anything like the way that lexical semanticists generally suppose, with facts about 'what the speaker/hearer knows *qua* speaker/hearer'. To see this, assume for the moment that semantics is purely denotational, so that the lexicon consists solely of a compendium of 'disquotational' truths like '"red" denotes (the property) *red*', '"dog" denotes (the property) *dog*', '"two" denotes (the number) *two*', and so on. Notice that it is perfectly possible for someone who believes in this sort of lexicon to hold that each such disquotational truth is analytic; that is, that the truth of each is guaranteed just by the semantics of English together with logic. But it would not follow from there being analyticities in *that* sense that grasping the truths of disquotation is a necessary condition for speaker/hearers to know the lexicon of their language. It is, for example, not plausible that understanding the word 'red' requires knowing what quotation is, or having the concept of denotation, or the concept of a property, and so on.

Nor would it follow that a lexicon that underwrites the notion of analyticity in this way would thereby underwrite synonymy, antonymy, and the rest. Suppose X and Y are absolute synonyms; and suppose that their lexical entries are, respectively, the disquotational principles '"X" denotes Xs' and '"Y" denotes Ys'; and suppose both of these disquotational principles are analytic. Still, nothing that is in (or that follows from) the lexicon says that 'X' and 'Y' are synonyms. That's because nothing that is in (or that follows from) the lexicon says that *being X* and *being Y* are the same properties. On the present assumptions, then, this lexicon supports a notion of analyticity but not of synonymy.¹¹

The moral, so far, is this: There is a tendency among lexical

¹¹ An 'informational' semantics—one that construes meaning in terms of nomic or causal world-mind connections—might have very much this character. By contrast, lexical semanticists simply take for granted that semantic facts are basically epistemological—that they are grounded in facts about what speaker/hearers know about their language. We think that they are ill advised to assume this. For discussions, see Fodor and Lepore (1992) and Fodor (1994).

semanticists to think that synonymy, antonymy, hyponymy, and the like must be bona fide linguistic phenomena because, after all, linguists have spent such a lot of time studying them; 'Thus I refute Quine,' Samuel Johnson might have said. But, in fact, their bona fides depends on very strong, and very dubious, semantical assumptions, for which, to our knowledge, no serious defense is thus far on offer. Methodology cannot constrain the lexicon to reconstruct analytic inference if, in fact, there is no such thing.

These remarks apply to lexical semanticists at large; but in fact JP's theory faces even greater problems, since many of the inferences that he thinks lexical entries determine (e.g., from 'wants a cigarette' to 'wants to smoke a cigarette'; see below) are 'defeasible' (1995: 46), hence not even strictly necessary. One might reasonably take the view that if the necessity of an inference does not imply the semantic relatedness of its lexical constituents, no weaker modality is likely to do so. JP's comments on these sorts of issues tend not to be illuminating. For example,

One fairly standard definition states that two expressions are synonymous if substituting one for the other in all contexts does not change the truth value of the sentence where the substitution is made . . . A somewhat weaker definition makes reference to the substitution relative to a specific context. For example in the context of carpentry, 'plank' and 'board' might be considered synonyms, but not necessarily in other domains . . . (1995: 23)

But neither notion of synonymy can be right since, on the one hand, no two words substitute, *salve veritate*, in every context (see Mates 1950); and, on the other hand, any two words substitute, *salve veritate*, in some context (e.g., 'elephants' and 'asteroids' substitute in the context 'are often bigger than a breadbox'). One could perhaps counter this objection if one had an account of how to individuate semantic 'domains'; but JP provides none.

JP does not address the problem that is implicit in making notions like 'defeasible inference' respectable as a basis for semantic theory, namely, the problem of distinguishing what the language tells one from what one knows about the world. 'My wife uses the subway every day' . . . is a near paraphrase of "My wife travels on the subway every day", an interpretation that is made possible by our knowledge of what the function of a subway is' (1995: 87). But if it is indeed

what we know about subways that makes this interpretation possible, then the conclusion should be that the two sentences are not 'near paraphrases' in any sense that semantic theory is concerned with. If lexical entries express *meanings*, then what's in the semantic lexicon is information about words, and *not* information about the world. But it is perfectly possible both to understand 'Sarah likes to use the subway' and to wonder what she likes to use it for.

JP's demand for a semantics of interlexical relations presupposes that the notion of analyticity is viable, hence that there is an answer to such questions as: 'What distinguishes linguistic knowledge from world knowledge?' and 'What distinguishes lexical entailment from mere necessity?' But we're impressed by the difficulty of deciding, in any principled way, which interlexical relations are meaning-constitutive; and, short of stipulating IRS, nothing so far rules it out that none are. Accordingly, in what follows we will assume that the failure of an atomistic, denotational lexicon to capture such relations as synonymy, hyponymy, and the rest is not a decisive argument against it.

Semantic Well-Formedness

JP writes, 'I will introduce a notion of *semanticity*, analogous to the view of grammaticality [*sic*] . . . but ranging over semantic expressions rather than syntactic structures. Semanticity refers to the semantic well-formedness of expressions in a grammar' (1995: 40).

No doubt, if there is such a property as semanticity, the lexicon should contribute to determining it. The trouble is that it's not clear what property it might be; JP provides no general characterization, and the few examples he offers are not transparent.

Consider pleonasms like ?'Mary kicked me with her foot' and ?'Mary buttered the toast with butter' (1995: 40). On any plausible account, what's wrong with these sentences is that the prepositional phrase adds nothing to what the rest of the sentence says, so a Gricean imperative ('Be informative') is violated. We suppose JP thinks that the locus of this redundancy is the lexical meaning of the verbs, the relevant consideration being that anyone who knows what 'to butter' means knows that buttering is done with butter. But, *pace* JP, this consideration does *not* argue unequivocally that the lexical entry for 'to

butter' specifies anything more (or anything other) than its denotation. Suppose that the dictionary says only that 'to butter' refers to buttering. Well, you can't learn this denotational fact unless you know *what buttering is* (viz., that it's covering with butter¹²), so anyone who knows what 'to butter' denotes *ipso facto* knows that 'Mary buttered the toast with butter' is redundant. It seems that this kind of 'unsemanticity' does not, after all, argue for 'a level of representation in the semantics, which operates according to its own set of constraints' (1995: 42). To assume that it does so is to beg the question whether 'no buttering without butter' belongs to the semantics of 'to butter' or to the metaphysics of buttering.

Here are some of JP's other examples of semantic ill-formedness (1995: 41):

- (1) (a) ?John began the dictionary.
- (b) ?Mary began the rock.

According to JP, these are 'semantically odd because of what we normally associate with the semantic possibilities of a noun such as "dictionary" and "rock"' (1995: 41). We're inclined to think, on the contrary, that there's nothing wrong with (1a) at all, and that, if there's anything wrong with (1b), it's that one can't imagine what it was that Mary began to do with(/to) the rock. Given a clue (e.g., she began, in either sense, to paint it), the perplexity vanishes and so does the intuition that something is awry with the sentence. (Cf., 'Having finished the oceans on Monday afternoon, God began the rocks on Tuesday morning.')

The trouble, in a nutshell, is this: if capturing semantic well-formedness is to be a constraint on representations of lexical meanings, hence on lexical entries, it has to turn out that what's wrong with a semantically ill-formed sentence is that *what it means* is defective, not just that it would be 'normally' or 'generally' hard to 'readily' interpret without contextual support. It is not clear that any of JP's examples have this property. Indeed, it is not clear that there *is* such a property as defectiveness of meaning as distinct from ungrammat-

¹² Actually, covering *x* with butter isn't sufficient for buttering *x* (though perhaps it's necessary; that depends on what 'covering' covers). That lexical entailments are rarely biconditional is the standard argument for not treating them as special cases of definitions (*pace* JP and both the interpretive and generative traditions in lexical semantics).

icality, necessary falsity, and any of a variety of kinds of pragmatic malfeasance. Barring more persuasive examples, or an independent argument for 'a level of representation in the semantics' whose rules would determine the semantic well-formedness conditions for expressions, we are inclined to doubt that failure to mark (la-b) and the like as defective militates very strongly against purely denotational lexical entries.

Distribution

It is a widely held view that the semantic properties of a lexical item 'give rise to' (1995: 19) its syntactic properties: syntactic distribution is somehow determined or explained by, or is anyhow predictable from, meaning; so a lexicon that represents only denotation would miss linguistically salient generalizations. In this spirit, Higginbotham (1994) remarks that 'the meanings of lexical items systematically infect grammar. For example . . . it is a condition of object preposing in derived nominal constructions in English that the object be in some sense "affected" in the events over which the nominal ranges: that is why one has . . . algebra's *discovery* (by the Arabs) [but] not *algebra's *knowledge* (by the Arabs)' (1994: 102).¹³ Similarly, according to JP, 'the diversity of complement types that a verb or other category may take is in large part also determined by the semantics of the complements themselves' (1995: 10).

It is, however, notoriously difficult to assess the claimed correlations between lexical semantics and syntactic distribution, because one is never told what the semantic representations themselves mean. What, exactly, is it for an object to be 'in some sense affected' in an event? This imprecision tends to undermine the supporting examples. So, for example, JP says that the difference between 'John is running' (therefore, John has run) and 'John is building a house' (*Therefore, John has built a house) 'is whether an action is homogeneous in nature or has a culmination of some sort' (1995: 16). Well, one might have thought that the process of *starting* to run has a

¹³ If you think your intuitions tell you that this is so, ask them whether algebra was more, or less, 'affected' by being discovered by the Arabs than the electric light was by being discovered by Edison. Our intuitions say such questions are plain nonsense.

culmination of *some* sort; namely, running. But 'John is starting to run → John has started to run' looks fine (cf., 'John is starting to build a house → John has started to build a house', etc.). Our point isn't that this is a counterexample to JP's generalization; rather, it's that the generalization is formulated so imprecisely that we *can't tell* whether it's a counterexample. JP's emendation to 'having a culmination' (viz., 'culminates in a changed state, i.e., . . . is an accomplishment' (1995: 16)) doesn't help much. Does it rule 'starting to run' in or out? By what criterion? After all, 'culminates in' is a *technical* term when it's used this way (as are, for example 'agent', 'patient', and 'theme' as they are used in the theory of theta-roles, to which the present points apply *mutatis mutandis*). Assuming 'culminates in' without explication in the language used to specify the semantics of lexical entries thus begs the key issue: whether predicates that are syntactically similar to 'build' do, in fact, share anything *semantic*.¹⁴

In fact, we know of no cases in the lexical semantics literature where the semantic end of putative semantics/syntax correlations has been made out with sufficient clarity to permit the claims to be evaluated. Another case in point is JP's example given in (2). Compare:

- (2) (a) The woman ate her meal.
- (b) The woman ate.
- (c) The dog devoured the cookie.
- (d) *The dog devoured.

Why can speakers drop the direct object of 'ate' but not of 'devoured'? The explanation, according to JP, is that 'while "eat" denotes an activity of unbounded duration (at least lexically), "devour", one might argue, denotes a transition . . . "devour" . . . carries a completive implicature that is absent from "eat"' (1995: 11).

There is some uncertainty about how much JP is claiming here; if (as seems likely; see 1995: 9–12) JP thinks that the fact that 'eat' is lexically represented as *denoting an unbounded activity* explains (i.e., that it provides a sufficient condition for) dropping the object, then

¹⁴ Another way to put this point is that, on the kind of theory JP advocates, lexical entries express senses, not just denotations. But the sense of 'culminates in' is at least as obscure as the sense of 'build', so it is hard to see how appealing to the one could much illuminate the other.

he must believe that there could not be a word that means what 'eat' does but does not allow its direct object to delete. This claim is remarkably strong and there is no reason to suppose that it is true. Notice that all of the following are bad, though unbounded activities are apparently involved in each: *'John stroked'; *'John ground (/his teeth)'; *'John pounded'; *'John smelled (/the salt air)'; and so on.

But, maybe JP thinks that unboundedness is just a *necessary* condition for the direct object to delete? Apparently not, since 'win', 'lose', and 'tie' are surely bounded and 'John won/lost/tied ← John won/lost/tied the race' seems fine. In any event, the usual problem crops up; we can't really evaluate the proposal because we don't really know what it is for an expression to 'denote an activity of unbounded duration (at least lexically)'. For example, nobody eats for ever; why doesn't that make 'eat' bounded? And why doesn't the 'change of state' from eating to not eating that occurs when one stops eating constitute a 'transition', just as much as the cessation of devouring that occurs when one stops devouring? If 'devour' carries a completive implication, why is 'She never finished devouring her meal' all right? And even supposing that 'devour' does carry a completive implication, so too, surely, does 'finish eating'. But 'The woman finished eating' is perfectly fine though *'The woman finished devouring' isn't.

The fact is, nobody knows what, if anything, it is for a word to denote a transition; so nobody knows whether 'devour' does. Similarly, we believe, for the other claims about the semantic determinants of syntactic distributions with which the lexical semantics literature abounds. What's worse, it's possible to wonder, in such cases, whether there really are two different things to correlate. Perhaps the intuition that 'devour', but not 'eat', is *semantically* completive is just a hypostatic misconstrual of one's *syntactic* knowledge that the first but not the second verb takes a mandatory direct object. One can't correlate a thing with itself; that is why Pooh never caught a Woozle. (See Milne 1957. For further discussion, with further examples, see Fodor 1998a, ch. 3.) It's hard to overemphasize the importance of this point, so we pause to rub it in once more. It's not in dispute that there is a syntactic class of object-deleting verbs (call it C), or that 'eat' belongs to C, or that 'devour' doesn't. What's in dispute is whether C is *semantically characterizable*, and, if so, what

the semantic characterization is. So far, the only suggestion on offer is that verbs in C all apply to 'bounded' events. But no independent account of what the boundedness of an event consists in is on offer, and (so far as we can tell) there are plenty of counterexamples to the claim that *being in C* entails *expressing a bounded event* are coextensive. *A fortiori*, we think, *no case at all* has been made that C is a semantic class; indeed, that its members share anything more interesting than some quirks of syntactic distribution. Corresponding remarks apply, *mutatis mutandis*, to every argument for the semantic explanation of a syntactic distribution that we've encountered in the lexical semantics literature.

But suppose that, despite the unconvincingness of the examples and the imprecision of the metalanguage that is used to describe them, it really is true, in God's eyes as it were, that (some or all of) lexical syntax is semantically driven. Would that show that there's something more to lexical representations than denotations? Not obviously. For, even if it is devouring's being a transition that explains why 'devour' has to have an object, why should knowledge of this fact count as part of what speaker/hearers know about 'devouring' rather than as part of what they know about *devouring*? *A priori*, it seems as reasonable that syntax should be driven by what people know about what a word denotes as that it should be driven by what they know about what the word means. Barring some principle for choosing between these two interpretations, we don't see how the putative influences of semantics on syntax could count in favor of a semantics of lexical meanings *rather than* denotations.

So much for the polemical background. Though one finds all these sorts of arguments in JP's book, they have a considerably older provenance, and JP's treatment of them is not more convincing than what earlier discussions offered. What makes JP's book interesting is his argument from the semantic generativity of the lexicon to the semantic complexity of lexical entries. We turn to that now.

Generativity

A word about method: JP develops a considerable apparatus in setting out his proposals about lexical generativity, and neither his

notation nor his exposition is strikingly perspicuous. Therefore, we will in most cases not reconstruct the technical details. Instead, we will consider several of JP's examples of how his theory applies to the analysis of linguistic data. We will try to show that these analyses don't work, but our main purpose is not to impugn them; rather, it is to illustrate the character of the theory and of its defects. We claim that JP's failures to get the data right are principled and show that there is something wrong with the *kind* of account of the lexicon he endorses. We aren't, in short, just quibbling about cases.

JP's discussion of generativity starts by invoking the distinction between 'polysemy' and mere accidental homonymy.¹⁵ Intuitively, the paradigm of polysemy is the kind of quasi-ambiguity exhibited by the 'believes' in 'John believes Mary', on the one hand, and 'John believes that 2 is less than 3', on the other. However, according to JP, polysemy can also be exhibited by expressions that, unlike 'believe' in this example, are syntactically homogeneous: for example, JP says that 'window' means *aperture* in 'He crawled through the window' but that it means *physical object* in 'The window is rotting'. Similarly, JP thinks 'good', though it has only one lexical entry, means something different in 'good knife' and 'good secretary'.

Many linguists have shared JP's sympathy with Weinreich's (1959) complaint that standard treatments of the semantic lexicon fail to reconstruct the difference between polysemy and mere homonymy. In effect, by postulating distinct lexical entries for each meaning of what is in fact a polysemous term, sense enumeration lexicons miss 'The fact that [the] two senses [of a polysemous expression] are logically related . . .' (Pustejovsky 1995: 37). However, on JP's view, the fact that 'believe' or 'window' (or 'use'; see below) often seem to take on different, but related, senses in different host expressions is itself just a special case of the perfectly general fact that 'words can assume a potentially infinite number of senses in context . . .' (1995: 105). Correspondingly, the theorist's goal is to capture this context-sensitivity of lexical meaning 'while limiting the number of senses actually stored in the lexicon' (1995: 105); that is, without proliferating lexical entries beyond necessity.

¹⁵ Merely terminological stipulation: We will take ambiguity to be the generic property of which polysemy and homonymy are species.

To see how this is to be accomplished, let's start with JP's treatment of the putative polysemy of 'bake'. *Prima facie*, 'bake' is polysemous between 'a creative activity' as in 'bake a cake' and a 'change-of-state predicate' as in 'bake a potato'. 'Intuitively, we would like to capture the fact that the former objects are prototypically brought about by the activity they are in composition with . . .' (1995: 123); that is, that 'certain objects come into being by virtue of an activity which can otherwise simply change the internal state of an object' (1995: 98).

The kind of 'sense enumeration' semantic analysis that JP wishes to reject might distinguish two lexical entries for 'bake': *CREATED* versus *HEATED UP* (or something of the sort). JP's account, by contrast, has a single lexical entry for 'bake' and distinguishes the semantics of 'bake a cake' from the semantics of 'bake a potato' on the basis of a difference in the lexical representations of the object nouns 'cake' and 'potato': 'we can derive both word senses of verbs like "bake" by putting some of the semantic weight on the NP. This view suggests that, in such cases, the verb itself is not polysemous. Rather, the creation sense of "bake" is contributed in part by the meaning of "a cake", by virtue of it being an artifact' (1995: 124).

This treatment is characteristic of JP's general approach; indeed, much of the book consists of applying it to a wide variety of examples. Notice, in particular, the idea that the sense of 'governing' expressions is inherited partly from their lexical entries, but partly also from the semantics of the expressions that they govern in a context. According to JP, this is true not just of the relation between verbs and their complements but also, for example, of the relation between adjectives and the nouns they modify (e.g., 'good' inherits *FOR CUTTING* from 'knife' in 'good knife'; 1995: 129). Notice also that pursuing this strategy places constraints on the lexical contents of the *governed* expressions. 'Use' can inherit *TO RIDE ON* from 'subway' in 'use the subway' only if the 'telos' (function) of a subway is part of the entry for 'subway'. Likewise, the creative sense of 'bake a cake' can emerge from the occurrence of *ARTIFACT* in the entry for 'cake' only if that entry specifies the information that cakes are artifacts; and so on. In effect, *the more context-sensitive the meanings of governing expressions are supposed to be, the richer must be the lexical entries for the expressions that they govern*. So, JP's standard tactic for arguing that there is a certain piece of semantic information in the lexical

entry for a noun is to show that there is a context in which that information is inherited by a verb to which the noun functions as a complement, or by an adjective that modifies the noun, and so on.

Since, in all these respects, JP's analysis of 'bake' is paradigmatic, the failure of the analysis is characteristic of the general inadequacy of the theory. We turn to considering this analysis at some length.

To begin with, suppose that JP is right and the reason that 'bake' in 'bake a cake' means *create* and 'bake' in 'bake a potato' means *warm up* is that cakes are artifacts and potatoes are 'natural kinds'. But would that explain why the two 'bakes' aren't merely homonyms? According to JP, '[Artifacts] such as cookies, cakes, and bread are typically baked. The process of baking, *modulo* such objects, is a creative activity, while relative to objects such as potatoes, carrots, and other natural kinds, it is simply a change of state predicate . . . to classify the verb "bake" as having both senses lexically specified is to miss the semantic generalization' (1995: 98; see also 47). Apparently, on this account, what makes the two 'bakes' not merely homonyms is that both express the same 'process of baking' which, in one case, constitutes a creative activity and, in the other case, merely effects a change of state.

But now: By what criterion do both kinds of baking count as the same process? What decides that the 'bake' in 'bake a cake' (thereby creating one) denotes the same activity as the 'bake' in 'bake a potato' (thereby merely heating one)? Whereas, by contrast, 'bank' is presumably homonymous because the banking in 'bank a check' counts as a *different* process from the banking in 'bank a plane'. Patently, this question why the two kinds of baking count as one process but the two kinds of banking don't is just the polysemy problem all over again; all that's happened is that it's been kicked upstairs from the semantics to the ontology. Whereas we used to worry about how to count senses, we are now invited to worry about how to count processes. Six of one, half a dozen of the other, as far as we can tell. It ought to be untendentious that you can't warrant your semantic intuitions by appeals to your metaphysical intuitions if the latter aren't independently grounded.

But, for the sake of the argument, let's put ontological qualms aside and go back to the linguistics.

If the 'creative' sense of 'bake' is determined by something that it

inherits from its direct object, and if 'bake' and 'cake' are themselves univocal, then 'bake a cake' must have *only* the creative reading. But, in fact, 'bake a cake' is ambiguous. To be sure, one can make a cake by baking it; but also one can do to a (pre-existent) cake just what one does to a (pre-existent) potato: put it in the oven and (non-creatively) bake it. Since 'bake a cake' is ambiguous and 'cake' is univocal, it must be that 'bake' isn't univocal after all, contrary to JP's analysis.

The ambiguity of 'bake a cake' shows that JP is wrong to think that 'bake' is univocal. But it does not show that he is wrong to think that 'cake' is lexically marked as an artifact. Indeed, at first blush it appears that *only* a lexically marked distinction between artifacts and 'natural kinds' could account for the fact that 'bake a potato', unlike 'bake a cake', appears to have only the noncreative reading; if 'bake' is ambiguous, both expressions should have both readings. So, maybe some of JP's program can be saved after all.

But, in fact, even assuming that the lexicon distinguishes artifacts from natural kinds is not enough to predict correctly the ambiguity/univocality of their hosts. Notice that although knives and trolley cars are artifacts, 'bake a knife' and 'bake a trolley car' resist a creative reading quite as much as 'bake a potato' does. But if 'bake a cake' is heard as creative because 'cake' is marked as denoting an artifact, then 'bake a trolley car' should be heard as creative too. Clearly, something has gone wrong.

In fact, there is a striking difference between JP's informal account of 'bake' and what he actually puts in the examples of lexical entries that he offers. According to the informal account, 'the creation sense of "bake" is contributed in part by the meaning of "a cake", by virtue of it being an artifact' (1995: 124). But his lexical entry for 'cake' specifies not that *cakes are artifacts*, but, in effect, that *cakes are made by baking* (1995: 123). (Compare JP's discussion of 'co-compositionality'.) Presumably, this is required in order to distinguish 'cake' from 'potato', on the one hand, and from 'knife', on the other: the entry for 'potato' says potatoes are a 'natural kind', and 'knife' is presumably lexically specified for some means of production other than (and exclusive of) baking (e.g., by grinding).¹⁶ In

¹⁶ This won't work, however, unless the semantics somehow provides the information that what is made by grinding is *ipso facto* not made by baking. How one gets such exclusion(/inclusion) relations to hold among semantic representations without,

short, JP has to explain why, although it's an artifact verb, 'knife' doesn't impose a creative interpretation on 'bake'; and this requires him to assume that not just ARTIFACT, but also MADE BY GRINDING (or whatever), belongs to the lexical entry for 'knife'. This means that you don't know what a term for an artifact means unless you know how that kind of artifact is made. If, like the present authors, you don't know how pencils are made, you don't know what 'pencil' means. But we do.

The right story is surely this: As far as the language is concerned, 'bake' is equivocal, and 'bake a potato' and 'bake a knife' are both ambiguous. What makes both sound funny on the 'creative' reading is something you know about the world, not something you know about words: namely, that you can't make a knife or a potato by baking it. If you didn't know this, you would hear the ambiguity, as indeed you do in 'John is baking something' or 'What is John baking?'¹⁷ In short, contrary to JP's analysis, 'bake' is lexically ambiguous and the semantics of 'bake NP' offers no argument that the lexicon contains ontological information.

But is the 'bake' case representative? We will look briefly at some of JP's other examples. We claim that JP offers no convincing examples where the meaning of a governing expression is modulated by the lexical content of the expressions that it governs. In fact, it's our view that this never happens.

Consider JP's treatment of 'begin'. The idea is that 'begin'(/'finish') picks up the 'telic role' of its direct object (1995: 115–17). This is to account for such inferences as (3a–c).

- (3) (a) *begin a book* → *begin to read a book*
- (b) *finish a cigarette* → *finish smoking a cigarette*
- (c) *begin a beer* → *begin drinking a beer*.

It's this sort of example that provides JP's primary evidence that information about function is part of the lexical entry for 'book', 'beer', and the like.

in effect, resorting to meaning postulates is a traditional, and unsolved, problem in the literature on lexical semantics.

¹⁷ Assuming that gapping is semantic, JP's theory requires that 'John baked the potatoes and Mary the cookies' means that Mary heated up the cookies, not that she made them. We don't think that prediction is true either. We suspect that gapping is semantic, and that it distinguishes polysemy from true homonymy; hence, *'He banked the plane and the check'.

On second thought, however, the analysis of 'begin' doesn't work in the general case since it isn't true in general that if what NP denotes has a function, then 'begin NP' means *begin to use NP to perform its function*. 'Begin a car' doesn't mean *begin to drive a car*; 'begin a thermometer' doesn't mean *begin to measure the temperature*; and so on. 'Enjoy', another of JP's salient examples of a verb that incorporates the telic role of its direct object (1995: 88), fails for the same sort of reason. JP notes cases like 'enjoyed the meal → enjoyed eating the meal'. But his account incorrectly predicts the well-formedness of *'enjoyed the doorknob', *'enjoyed the federal government', *'enjoyed the carpet tack', all of which are bad. The first and last of these examples are especially problematic for JP's account since 'enjoyed using the doorknob(/carpet tack)' is fine.

Indeed, 'use' may seem more suitable for JP's purpose than 'begin' or 'enjoy', since if X uses Y, where Y is something that has a (conventional/typical) use, then the invited inference is that X uses Y for what Y is used for. (Even so, there's something odd about 'used a glass of beer', 'used a meal', and the like.) In fact, however, 'use' raises a thorny problem for JP's account since it's perfectly possible to use things that don't have uses: a rock to break a window, snow to make a snowball. ('Enjoy' works this way too since one can perfectly well enjoy things that have no function (the clement weather; dancing the eightsome reel; etc.)) These examples raise a serious question about which JP has little to say: What happens if a verb makes a demand on an argument that the lexical entry of the argument doesn't satisfy?

JP's view may be that, if the governing expression demands X in the lexical entry for a governed expression that is not marked for X, the resulting phrase is interpreted as existentially generalized in respect of X. We infer this from JP's treatment of 'Mary believes John' (1995: 120–2). According to JP, 'believes' wants a proposition in the lexical entry for its direct object, which, however, the lexical entry for 'John' presumably does not contain. Though the mechanism by which it's achieved is unclear to us, JP's solution is that the interpretation of 'Mary believes John' contains an existential quantifier over the required propositional argument: something like *Mary believes what John communicated*.¹⁸ Assuming that 'believe John' constitutes

¹⁸ In particular, according to JP, the semantic representation of *Mary believes John* is the existential generalization of the P2 in BELIEVE (P2 (JOHN)) (MARY) (1995:

a precedent for 'use a rock', then 'John used a rock' should come out meaning something like *John used a rock to perform some function*.

If, however, that is the treatment that JP has in mind, it's certainly untenable. The problem is that, given interpretive mechanisms that strong, it becomes unclear how a combination of a verb with an NP argument could ever fail to be interpretable. For example, JP's own analysis of 'begin a rock' is undermined; it should be (but isn't) heard as meaning *begin to use a rock for something*. In like wise, why doesn't 'John asserted a rock' mean *For some rock-involving proposition P, John asserted P*? Conversely, why doesn't 'Bill used that John is tall' mean *For some function F, Bill used the proposition that John is tall to perform F* (compare 'Bill used John's being tall to illustrate the effect of diet on growth')? In order to argue for telic roles in the lexical entry for NP arguments, what's really needed is a verb that picks the telic role when there is one and that yields unsemanticity when there isn't one. But, as far as we know, there is no such verb.¹⁹

Let's summarize the argument so far. JP's general strategy is to explain the apparent polysemy of governing expressions by appealing to the semantic heterogeneity of the objects they govern: if X is a governing expression that is *prima facie* A/B polysemous, assume that the expressions that it governs are lexically cross-classified as being either A or B. We know of no cases where this strategy works. For one thing, it predicts that an A/B polysemous expression should be univocal in the A direction when it governs a univocal A-expression,

122). The problem is, where does the propositional constant P2 in this formula come from? Surely not from the lexical entry for *John*, since *John* doesn't mean anything propositional. As JP himself says, 'Unlike the case of . . . the book . . . the [proposition] type required by selection [is not part of] the [lexical] interpretation of the complement' (1995: 121). As far as we can tell, JP never suggests an answer to this question. (*About* raises a structurally analogous problem, since it seems to govern an event in *a book about Vietnam* (i.e., *about the Vietnam War*) but not in *a book about Nixon*. Where does this event come from? JP says only that *about* 'covers' objects of two different ontological types in the two cases—which, we suppose, is just to say that it's polysemous.)

¹⁹ On pp. 137 ff., JP provides an (unconvincing) explanation of *John asked me (/wondered) the temperature*. But he doesn't consider why, given that *John wondered what the temperature is* is well formed, the statement *John wondered the temperature* doesn't have a reading on which *the temperature* is 'coerced' to an interrogative proposition.

whereas, in all the cases we can think of, A/B polysemous expressions are A/B ambiguous in both A and B contexts (cf. 'bake a cake'). For another thing, the claim that governing expressions inherit semantic content from and only from the lexical entries of the expressions they govern fails in both directions: 'begin' inherits a 'telos' from 'a cigarette' but not from 'a car', and the propositional construal of 'believes John' can't be derived from the lexical entry for 'John'. And so on.

We conclude:

1. Apparent polysemy is generally real; the reason 'bake' seems to be lexically ambiguous is that it is. This is compatible with a denotational semantics on the assumption that baking (creative) and baking (warming up) are different processes. JP provides no grounds for doubting this assumption.

2. There is no evidence that the meaning of governing expressions is ever modulated by the semantics of the expressions that they govern. For all the arguments show so far, 'bake' behaves the same way in 'bake a cake' and in 'bake a potato' (it's ambiguous in both); 'enjoy' means the same thing in 'enjoy a cigarette' and in 'enjoy the sunset' (it means *enjoy*); 'good' means the same thing in 'good knife' and in 'good car' (it means *good*); and so on. If 'bank' means something different in 'bank a check' and 'bank a plane', then 'bank' is not polysemous but homonymous, and the context effects are not modulation but selection.

Compositionality and Logical Form

We start this section by distinguishing two kinds of issues that JP's treatment generally runs together: on the one hand, questions about how *complex* lexical entries are; and, on the other hand, questions about the *generativity* of the lexicon. As we have seen, JP typically argues for lexical complexity by claiming that it is needed to account for generativity. 'Cake' must contain ARTIFACT in order to account for the polysemy of 'bake'; 'subway' and 'telephone' must each contain a 'telos' in order to account for the polysemy of 'use'; and so on. However, in principle, the issues are dissociable. Someone who agrees with us that JP's arguments for lexical complexity are unconvincing

and who shares our suspicion that content is just denotation could wonder, nonetheless, whether a denotational lexicon might not be generative. So let's turn to this.

Suppose, for the sake of the argument, that the lexical entry for 'cake' contains *MADE BY BAKING*. What happens next? Actually, JP gives two different accounts, though he doesn't acknowledge that they are distinct. First: the lexicon is generative; 'cake' contributes *MADE BY BAKING* to the meaning-in-context of 'bake', thereby bestowing a 'creative' sense on the verb.

Thus, we can derive both word senses of verbs like 'bake' by putting some of the semantic weight on the NP. This view suggests that, in such cases, the verb itself is not polysemous [i.e., it has only one lexical entry]. Rather, the creation sense of 'bake' is contributed in part by the meaning of 'a cake' . . . The verb appears polysemous because certain complements . . . add to the basic meaning by co-composition. (1995: 124)

Second:

The semantics for the VP 'bake a cake' results from several operations . . . The operation of co-composition results in a qualia structure for the VP that reflects aspects of both constituents. (1995: 124; our emphasis)

The result of co-composition is a semantic representation at the VP level that is identical in structure to the lexical form for a creation verb such as 'build'. (1995: 125; our emphasis)

According to this second account, the lexicon isn't generative after all; 'cake' never contributes anything, at any level, to the representation of 'bake'. Rather, 'cake' contributes its lexical content (atomistic or otherwise) to the interpretation of the VP 'bake a cake' (as does, *mutatis mutandis*, 'bake').

Our polemical reason for insisting on the difference between these two accounts is that it's news if the *lexicons* of natural languages are generative; lots of people (ourselves included) think they are more or less lists. But it's no news that VPs are generative,²⁰ or that their semantics must somehow integrate the meaning of a verb with the

²⁰ We aren't denying that JP's two accounts are compatible. Perhaps he thinks that 'cake' contributes to the interpretation of 'bake a cake' by modulating the meaning of 'bake'. Our point will be that evidence that 'cake' contributes to the meaning of 'bake a cake' is neutral as to whether the lexicon is generative.

meanings of its argument. It is, however, only the latter sorts of conclusion that JP's treatment of 'bake' actually argues for.

We have a second reason for stressing the distinction between the view that verbs receive content from their arguments and the view that VPs do. Namely, that though we're very dubious that the meaning of a word is ever a function of its context, we have no doubt at all that (barring idioms) the meaning of a phrase is typically a function of the meanings of its lexical constituents. All semanticists, whatever they may think about the generativity of the lexicon, have to face the problem of the compositionality of lexically complex expressions: the problem, that is, of saying how lexical semantics contributes to determining the semantic interpretation of phrases. And, though we take the question to be largely moot, we are inclined to think that the exigencies of the compositionality problem really do require that lexical entries cannot specify only denotations; not even assuming that denotation is all that there is to content.

Consider the phrase 'want a beer'; and assume—what, to be sure, is tendentious—that the right interpretation of this phrase is *want to have a beer*,²¹ whereas, by contrast, the right interpretation of 'drink a beer' is just *drink a beer*. How does a theory that derives the meanings of phrases from the meanings of their lexical constituents ensure this difference of interpretation?

One way would be to agree with 'Sense Enumeration' lexicons that there are two 'wants', differing somehow in meaning, one of which takes an infinitival complement and one of which takes an NP. But here we agree with JP; this treatment is too unrevealing to be plausible. In particular, it misses the *prima facie* equivalence of 'wants a beer' and 'wants to have a beer'; and it fails to explain why, if Bill wants to drink a beer and Mary wants a beer, then they both want something and they both want to have something;²² and it fails to explain why the two 'wants' are in complementary distribution. So, what's the right way to handle 'want'?²³ Roughly, we think, like this:

²¹ On JP's account, of course, this is *not* the right interpretation; it is part of his story that 'want' is generative and that 'want a beer' means *want to drink a beer*. We'll return to this presently.

²² We think you should be prepared to admit this even if you doubt there is something (viz., to have a beer) that Bill and Mary both want or both want to have.

²³ We are considering only whether 'want NP' is polysemous and ignoring questions that are raised by its intentionality (viz., by the opacity of the NP position to

'Want' denotes the relation that holds between x and y when and only when x is a creature & y is a state of affairs (or whatever kind of thing infinitive expressions like 'to have a beer' denote), and x wants y .²⁴ The way to avoid lexical polysemy of 'want' is to maintain that it denotes this relation in both 'wants a beer' and 'wants to have a beer'; that is, it denotes this relation whether its complement is an infinitive or an NP.

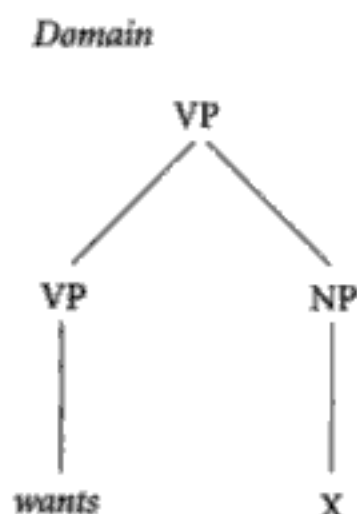
Of course, this proposal faces an immediate problem: a beer isn't a 'state of affairs' (i.e., it isn't the sort of thing that infinitival complements denote). So, then, how could 'want a beer' and 'want to have a beer' express the same relation? This is just the polysemy problem over again; how can the semantics ensure that expressions of the form 'want NP' get the same semantic value (viz., the same denotation) as corresponding expressions of the form 'want INF' without assuming two lexical entries for 'want'? Or, to put the same question in slightly different terms, how does a compositional semantics operate to assign the interpretation '*wants to have NP*' to expressions of the form 'wants NP'?

Here is our proposal: A lexical entry is allowed to be complex. If it is, it specifies (a) a meaning (viz., content, viz., denotation) and (b) a rule of composition that contributes to determining the *logical form* of the phrases of which the item is a constituent.

A (very) schematic derivation of the interpretation of expressions of the form *wants X* (where X is a lexical item) should serve to introduce the general idea. We assume that the semantic interpretation proceeds, node by node, from the bottom to the top of a (surface) syntactic tree. In the present case, the crucial steps are as follows:

the existential generalization and substitution of coreferentials). If you like, assume that *wants* expresses an 'intentional relation', and plug in your favorite account of them.

²⁴ Actually, the talk of creatures, states of affairs, and the like is entirely heuristic. That is because, on the present view, the lexicon is not required to specify a *sense* for 'want' but only required to say what 'want' denotes. Perhaps, if x wants y , then, necessarily, x is a creature and y is a state of affairs. But a purely denotational lexicon is not required to say that this is so, any more than it is required to say that dogs are animals or that 2 is prime. A denotational theory thus avoids the commitments to unanalyzed semantic entities and properties ('agents', 'patients', 'events', 'transitions', 'being in some sense affected by', etc.) by which, as we remarked above, standard views of lexical semantics are plagued.



Stage 1

Input: (*wants . . . X NP*)

Operation: Assigns lexically specified semantic interpretations.

Output: Interpretations of (viz., assignments of denotations to) the lexical nodes: Assigns to the V node the set of ordered pairs $\langle y, x \rangle$ such that y wants x , and assigns to the NP node the lexically specified denotation of X .

Stage 2

Input: The domain tree with the lexical nodes interpreted as per stage 1.

Operation: Interprets the node VP.

Output: VP is assigned ' $\{y: y \text{ wants to have } F(X)\}$ ', where $F(X)$ designates the interpretation that X receives in stage 1.²⁵

Note: We assume that the operation in stage 2 is driven by a composition rule that is part of the lexical entry for 'want': namely, if the constituents of VP_i are ('want X NP'), then the interpretation of VP_i is *wants to have* $F(X)$.

Several comments, in no particular order.

1. The derivation of 'wants to drink a beer' assumes that the surface syntax of that sentence is not relevantly different from its logical form. The composition rule assigns to the VP node the set ' $\{y: y \text{ wants to drink a beer}\}$ '. In this case, the composition rule is pre-

²⁵ If the sentence under interpretation is of the form ' NP_1 wants X ', a later cycle will identify with the denotation assigned to NP_1 .

sumably not lexically governed; this is the unmarked treatment for phrases consisting of a verb with an infinitive complement.

2. Notice that, according to this proposal, 'want' never means (denotes) anything except a relation between a creature and a state of affairs; not even in 'wants a beer' (i.e., not even in expressions where the surface complement fails to denote a state of affairs). Therefore, 'want' is not polysemous; its *content* (unlike the logical form of its hosts) is absolutely context-invariant. We're as far from a generative lexicon as it is possible to get. The cost of this univocality is complex lexical entries, which determine not only the content of an item but the host's logical form.

3. The proposed mechanism does the same sort of job that JP's notion of 'type coercion' is designed to do. There are, however, differences other than the fundamental one that distinguish a lexical semantics of complex senses from an atomistic lexical semantics of denotations. For example, it's part of what we take to be the context-insensitivity of *want* that it always introduces the same 'light verb' into the VP it governs. This repeats the remark we made in note 19: we assume that 'wants a pretzel' means *wants to have a pretzel*, not *wants to eat a pretzel*. The meaning of 'want' doesn't decide what one can want a pretzel for.

Indeed, we think, exactly contrary to JP, that it is this consideration that distinguishes expressions like 'want', which introduce light elements, from expressions like 'use', 'enjoy', and 'begin', which don't. In our view, the semantic interpretation of 'uses NP' is the set '{ y : y uses NP}'. Part of the evidence for this is that there is no relation (other, of course, than using) that y has to have to the NPs that y uses. Contrast the necessary truth that every y that wants NP *ipso facto* wants to have NP.

4. There is a variety of lexically governed effects on logical form other than light verb introduction. For example, we think it plausible that 'good' introduces a quantifier into the interpretation of 'good NP'; roughly, a good NP is one that is good for *whatever it is* that NPs are supposed to be good for (cf., Ziff 1960). Notice that this treatment makes the meaning of 'good' context-insensitive; 'good' quantifies over the function of the NP it modifies, and the way it does so is independent of which NP that is. Since the meaning of 'good' is context-independent, the lexical entry for the NP need not specify a

'telos'; so the semantics of 'good knife' provides no argument that 'knife' has a definition that includes its function. (This is just as well since, as JP himself remarks (1995: 43), 'good children' and 'good weather' are perfectly fine, though neither children nor the weather have functions. See above for the corresponding point about 'enjoy' and 'use'.)

5. Our discussion has not assumed that there is a *level* of logical form at which, for example, 'want a beer' is represented as WANT TO HAVE A BEER. On our account, all that happens is that wanting to have a beer (material mode) is assigned as the denotation of the expression 'wants a beer'. Our treatment is, however, compatible with positing an explicit level of logical syntax should there prove to be any reason to do so; we have no views on the matter for present purposes.

We close with two pending questions; the first strikes us as not very urgent, but the second is vital.

First point. What about polysemy? We don't have anything general to say about polysemy beyond the suggestion, implicit in the preceding, that where it is sensitive to the syntactic structure of the context, polysemy belongs not to the theory of content but to the theory of logical form. That leaves many residual cases like 'lamb' (*meat* v. *animal*), 'window' (*the opening* v. *what fills the opening*), 'newspaper' (*the thing that is read* v. *the organization that publishes it*).

We suspect that there is nothing that is both interesting and general to say about such cases; the meanings of words can partially overlap in all sorts of ways, so there are all sorts of ways in which polysemous terms can differ from mere homonyms. Nothing in the literature convinces us that there are powerful generalizations to state. Surprisingly, JP apparently shares this view so far as the polysemy of *nouns* is concerned (see 1995: 90–5). In cases like 'lamb' and 'window', JP's account does exactly what sense enumeration lexicons do: it has branching lexical entries that allow one to say, for example, that 'lamb' always means 'physical object', but doesn't always mean *animal* or always mean *food*. This is really unavoidable, given JP's overall architecture; since governing expressions get their ambiguity from what they govern, the governed expressions have to get theirs from the lexicon. Of course, as JP himself points out when discussing sense

enumeration lexicons, appealing to branching entries to distinguish polysemy from homonymy is a merely notational solution. It 'accounts for the data, but in a *post hoc* fashion, without making any predictions as to whether a particular datum should be possible or not' (1995: 42). 'Bank' *river* versus 'bank' *building* is presumably homonymy rather than polysemy, but both 'banks', like both 'lambs', mean (hence are 'branches of' the representation that corresponds to) *physical object*.²⁶ Nor is it clear why, if nouns can be really (viz., lexically) polysemous, verbs can't be too.

But be all that as it may; suppose that there could, after all, be a theory of polysemy that's worth having. Still, it strikes us as patent that it wouldn't be of the kind of theory that JP has in mind. JP has hold of the wrong end of the stick; he's got the order of explanation backwards. *Pace* JP, the question what a word refers to is prior to the question whether it is polysemous, not the other way around.

Suppose it's right that 'lamb' is polysemous between the animal and the meat. Surely that's because lamb-the-meat comes from lamb-the-animal. Surely there just *couldn't be* a word that's polysemous between *lamb-the-animal* and (say) *beef-the-meat*? Or between *lamb-the-animal* and *succotash-the-mixed-vegetable*? That there couldn't may itself sound like a deep fact of lexical semantics. But no; it's just the truism that, the less one can see what the relation between X and Y might be, the more one is likely to think of an expression that is X/Y ambiguous as homonymous rather than polysemous. Opportunities for polysemy arise from *how things are in the world* (or, anyhow, from how we take them to be). Correspondingly, if there's a productive generalization about the lamb/lamb kind of polysemy, it goes something like this: 'if X-stuff is made of Ys, there can be a lexical item that is X/Y polysemous'. Notice that this observation belongs to referential semantics, if it belongs anywhere.

This view of polysemy is not in the least novel. Indeed, it's built into traditional treatments of 'figures of speech' and the like. So synecdoche ('head' / 'head of the table') is, according to the usual dictionary definition, a figure in which a term for X is used as a term for

²⁶ In particular, we're not clear about the function of JP's dot operator (see 1995: 92 ff.), but we suspect that it has the effect of allowing components of lexical entries to combine freely, in the way that features do. If that is the right reading, then JP's lexical entries are even less constrained classical SELs like Katz and Fodor's (1963).

part of X, where *part of* is taken to be a *real relation*²⁷ among things in the world (in particular, among Xs and their parts). Only a linguist who is suffering from an acute case of semantics-in-the-head could miss this truism.

Second point. Prima facie, our notion of semantic coercion is more constrained than JP's. For example: according to us, it can't turn out that the content of a lexical expression ever depends on the content of its host; it isn't allowed that 'wants NP' sometimes means *wants to eat NP* and sometimes means *wants to drink NP*.

Nevertheless, we do allow that the logical role of an expression can be determined by the lexical entries of the forms that govern it; 'beer' is the logical object of 'drink' in 'wants to drink a beer', but it's the logical object of 'have' in 'wants a beer'. This departs from the most rigorous notion of compositionality, according to which each constituent contributes its content *and only its content* to its hosts, and the effect of a constituent on its hosts is absolutely context-independent.

The entirely rigorous notion of compositionality seems to us almost certainly not attainable. For example, 'want' and the like to one side, it is a truism that *logical* vocabulary is typically defined 'in use' (i.e., in a grammatical context), so that the lexical entry for (as it might be) 'the' determines the logical role of the NP in 'the NP' and it does so in a way that is specific to 'the'; presumably the logical role of the NP in 'an NP' is quite different. Indeed, it is arguable that the lexical content of 'the' is exhausted by what it says about the logical form of 'the NP'.²⁸

But if pure compositionality can't be had, it also can't be that just anything goes. What we've just said is tantamount to: *there can be*

²⁷ Hence not a 'Cambridge' relation. The putative problem of polysemy belongs to the same bundle as the problem of distinguishing real relations from the kind that philosophers make up. It's an open question whether either problem has a solution.

²⁸ By contrast, the lexical entry for 'and' is, of course, not exhausted by the compositional rule it gives; it also must supply a denotation, so as to distinguish between sentences of the logical form 'P & Q' and sentences of (e.g.) the logical form 'P ∨ Q', 'P → Q', and so forth. (Likewise, *mutatis mutandis*, with the entries for 'all' and 'some'.) If this is right, then there are perhaps four kinds of lexical entries: ones that specify only a composition rule ('a', 'the'), ones that specify only a denotation ('Tom', 'eat'), ones that specify both ('want', 'believe', 'good'), and ones that specify neither (expletives and the 'do' of *do*-support).

context effects on what a lexical item contributes to logical form but not on what it contributes to content. But since we don't know how to be rigorous about the form/content distinction, we're not persuaded that the notion of coercion that we've sketched actually succeeds in constraining the ways that a constituent can affect its hosts. Which is to say that, though we're sure that language is compositional, we don't know what the claim that it is amounts to. That, and not polysemy and the like, seems to us to be the main problem of lexical semantics.

Impossible Words?

THE idea that quotidian, middle-level concepts typically have internal structure—definitional, statistical, or whatever—plays a central role in practically every current approach to cognition. Correspondingly, the idea that words that express quotidian, middle-level concepts have complex representations ‘at the semantic level’ is recurrent in linguistics; it is the defining thesis of what is often called ‘lexical semantics’, and it unites the generative and interpretive traditions of grammatical analysis. Hale and Keyser (1993) (HK) have endorsed a version of lexical decomposition according to which ‘denominal’ verbs are typically derived from phrases containing the corresponding nouns: ‘sing_{vr}’ is supposed to come from something like DO A SONG; ‘saddle_{vr}’ is supposed to come from something like PUT A SADDLE ON; ‘shelve_{vr}’ is supposed to come from something like PUT ON A SHELF; and so forth.¹ Their case for these claims revives a form of argument, the ‘impossible-word’ argument, which has for some time been in eclipse. In this article we will claim that, whatever the right account of lexical decomposition eventually turns out to be, impossible-word arguments are infirm and should be discounted. We will use HK’s paper as a text for this critique.

¹ Cited forms are in single quotation marks and expressions in italics either stand for canonical names of meanings or are emphasized expressions. We use capital letters for canonical names of semantic representations. (So, according to the compositional view of causatives, ‘kill’ derives from CAUSE TO DIE and means *cause to die*.) However, for convenience, we will sometimes not distinguish between a semantic representation and the corresponding surface phrase; for example, we will say that ‘kill’ is derived from ‘cause to die’ to abbreviate the view that ‘kill’ and ‘cause to die’ are both derived from CAUSE TO DIE. Corner quotes (‘ ’) enclose variables that range over italicized formulas.

HK's arguments for the derivational analysis of denominal verbs are all variations on the same general theme: the derivational account predicts/explains the intuitive impossibility of certain verbs by showing that the syntactic processes that would be required to derive them are prohibited by independently well-confirmed grammatical constraints. We think there are principled objections to this form of argument. So, rather than dispute over details, we will take HK's word about what grammatical constraints are independently well confirmed; if they say that a derivation is blocked, we will assume that indeed it is. (We remark in passing, however, that a number of our colleagues have advised us against conceding (e.g.) that structures like (9) (see below) violate the empty category principle.)

Since we think that the logic of impossible-word arguments—hence their limitations—has not been made clear in the literature, we start by trying to be explicit about how such arguments are supposed to run. Three points about them will be assumed throughout our discussion.

First, the *conclusion* of an impossible-word argument is that there is plausible evidence for a grammatical rule/operation of *lexicalization*. (For exposition, we assume that lexicalization is a transformation.) The defining property of lexicalization is that it derives a surface lexical item from an underlying semantic representation. In the cases of interest, (a) the presumed underlying representation has internal constituent structure (in effect, it is a phrase), and (b) the surface item is more or less synonymous with the semantic representation from which it is derived.²

Second, the basic form of argument for lexicalization is that it satisfies independently motivated constraints on syntactic rules. Since it does, a grammar that fails to recognize lexicalization thereby misses a certain generalization, namely, that the formation rules for lexical

² There is considerable unclarity in the lexical semantics literature about just what meaning relations are required to hold between a lexical item and the semantic representation from which it is derived. Some claim, for example, that such derivations capture only the core or paradigmatic meaning(s) of surface lexical items. We will be neutral on this issue. We are concerned with the evidential status of the claim that surface lexical items *ever* derive from underlying phrases, whatever semantic relations such derivations are supposed to preserve.

items are a species of the formation rules for overtly complex linguistic structures like phrases.

Third, the evidence for the claim that lexicalization satisfies independently motivated grammatical constraints is that lexicalizations that violate them produce intuitively 'impossible words'. Impossible words correspond not to *de facto* gaps in the lexicon but to expressions that could not occur in a language *modulo* the grammatical universals.³ In fact, all of HK's examples are of the following form (1993: 60–4): they present a sentence that contains a neologistic verb, together with a paraphrase; and what is explained by appeal to the putative independently motivated grammatical constraints is why the former could not be derived from the latter (more precisely, why the former could not be derived from the representation that directly underlies the latter). HK say that the existence of constraints that block these derivations explains why 'English simply does not have verbs . . . that . . . have meanings corresponding more or less to the . . . paraphrases given here' (1993: 59–60).

We especially wish to emphasize how this sort of argument fits into the discussion about lexical decomposition (i.e., about whether surface lexical items sometimes have complex semantic representations underlying them). From that perspective, the question of interest is not whether there are impossible words; it is not even how to explain why impossible words are impossible. Instead, it is this: *Does the putative explanation of why they are impossible offer an argument for lexicalization?* That is, does it offer an argument that some surface lexical items are derived from underlying phrasal representations? We will argue that it doesn't and couldn't.

Prima facie, there are two kinds of ways in which constraints on possible lexical items might be predicted from general properties of grammatical derivations. Roughly, it might turn out that the deriva-

³ We are assuming that the constraints in question are universal. A weaker view might relativize the impossibility of lexical items to particular languages (e.g., *w* is impossible in *English* if its derivation would require rules or structures that English grammar happens not to acknowledge). As we understand them, impossible-word arguments generally have the stronger notion of constraint in mind; but the considerations we will raise do not depend on the distinction.

tion of a lexical item is blocked because it would violate some independently motivated constraint on the *kinds of grammatical rules* there can be, or because it violates some independently motivated constraint on the *kinds of structures* that grammatical rules are allowed to generate.⁴ (Though they do not discuss this difference, HK offer examples of both kinds.) It will be convenient for our exposition to consider these two kinds of blocked derivations separately since there are slightly different reasons why neither can provide evidence for lexicalization. But it's well to keep the bottom line in mind: we claim that neither of these kinds of impossible-word arguments can work. And since, as far as we can tell, these two are all the kinds of impossible-word arguments there could be, we also claim that impossible-word arguments are infirm in principle: *No impossible-word argument could be evidence for lexicalization.*

Impossible Words

In case 1 a derivation is blocked because a rule (a movement rule in all the examples that HK give) that the derivation depends on is impossible. For example, HK claim that there could not be a verb 'to cow' such that (1) is well formed and means *a cow had a calf*.

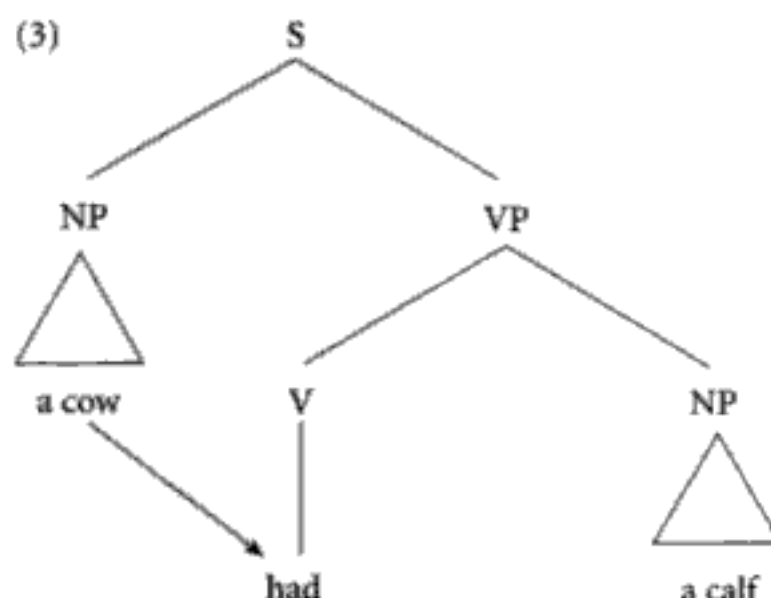
(1) *It cowed a calf.

They explain this as follows. Since, by assumption, lexical items typically derive from the structures that underlie their phrasal synonyms, sentence (1) would presumably come from the same structure that underlies sentence (2).

(2) A cow had a calf.

The structure of (2) is presumably (3), and the derivation would crucially depend on a transformation of subject lowering (followed by insertion of dummy *it* in the surface subject position).

⁴ This corresponds to the familiar distinction between derivations that block because they do not terminate and derivations that block because they generate structures that are subsequently filtered.

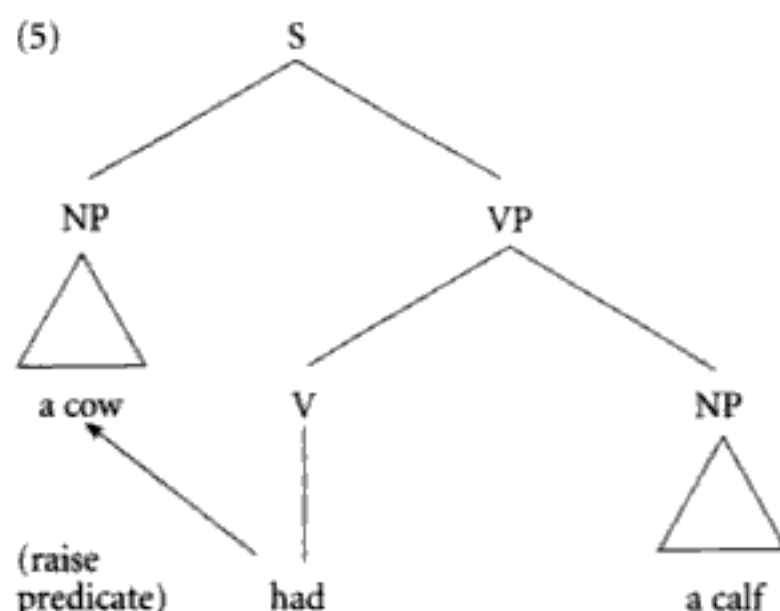


But, as HK remark, 'It is well known that a subject . . . cannot incorporate into the verb that heads its predicate' (1993: 60). Since subject lowering is an illicit operation, the putative derivation (3) is impossible; hence the intuition that (1) is likewise impossible.

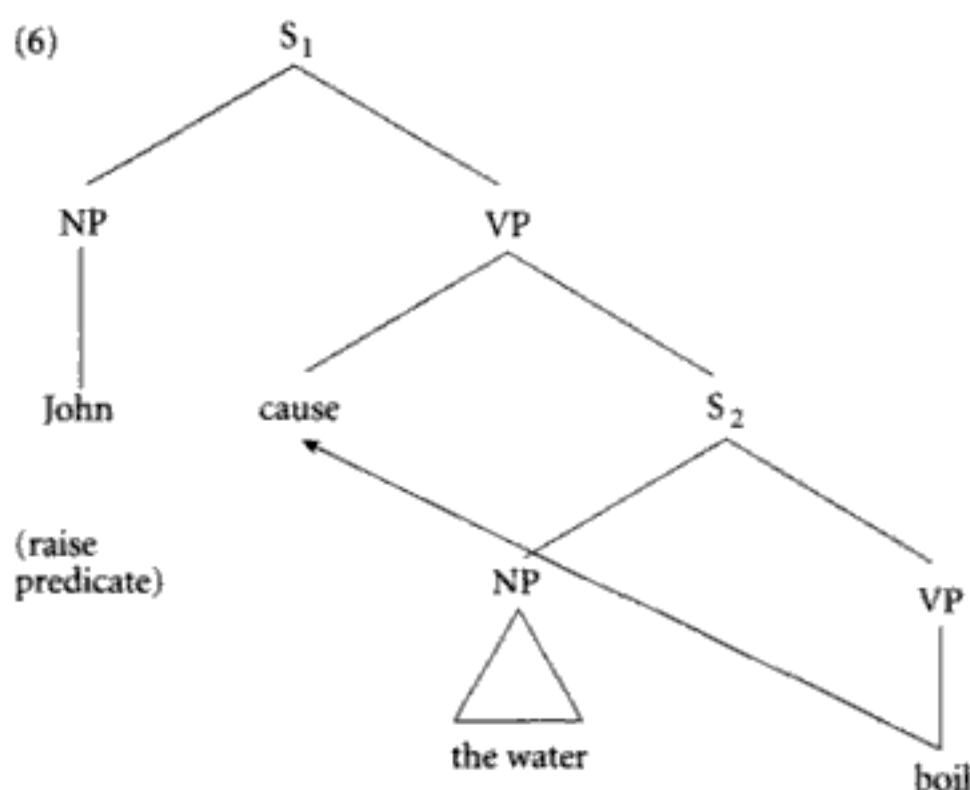
A preliminary comment: We think that this account is not plausible even independently of the general issues about lexical decomposition. In effect, HK trace the nonexistence of (1) to a (presumably universal) prohibition against transformations that lower subjects into predicates. Notice, however, that (4) is also impossible on the reading (1).

(4) *A cow did a calf.

However, the derivation of (4) requires, not lowering the subject NP, but raising the predicate (followed by insertion of dummy *do* in the surface verb position). Since (4) is impossible, the derivation (5) presumably has to be blocked.



But presumably (4) cannot be blocked by a prohibition against raising predicates, since such raising is required to derive (e.g.) causative transitives from their intransitive counterparts, as in (6). (6) is, in fact, a kind of derivation of which HK approve (see 1993: 79).



The symmetry between (1) and (4) suggests that the right explanation of the impossibility of the former is not that subject lowering is illegal. Rather, it is that deriving either (1) or (4) from (3) would

involve lexicalizing an expression that itself is not a well-formed constituent, namely, [cow did]. That lexicalizing a nonconstituent is illegal would presumably follow just from the assumption that lexicalization is a transformation; in any case, it is widely taken for granted in the generative semantics literature that only constituents can be lexicalized.

But we do not insist on this analysis. We are about to present an argument that neither examples where derivations block because the rules that they employ are illegal, nor examples where derivations block because the structures that they generate are illegal, can provide evidence for lexicalization. Here is our argument against the first disjunct.

There must be something wrong with HK's account of cases like (1) since, even if it did explain why there couldn't be a *derived verb* 'to cow' with the paraphrase (2), it does not explain why there couldn't be a *primitive*, underived verb 'to cow' with the paraphrase (2). As far as we can tell, this sort of point applies to any attempt to explain why a word is impossible by reference to the impossibility of a certain transformation. By definition, impossible-word arguments purport to explain intuitions of the form 'there couldn't be a word *w* that means "*e*"' by showing that [*e*] couldn't be a derivational source for *w*. But, on the face of it, that doesn't show that there couldn't be a word that means *e*; the most it could show is that if there is a word that means *e*, then it must be primitive.⁵ We assume, along with HK, that the intuition about (1) is that it is *impossible*—and not just that *if* it is possible, then it is underived. (We do not suppose that anyone, except perhaps linguists, has intuitions of the latter kind.) So we claim that HK have not explained the intuition that 'to cow' is impossible.

We repeat that this sort of objection holds in any case where a derivation is said to be blocked by the lack of transformations that would construct an appropriate domain for lexicalization—as, for example, the lack of a subject-lowering transformation is supposed to explain why there is no 'a cow had' to provide a domain for the introduction

⁵ We put it this way because, strictly speaking, a demonstration that [*e*] is not well formed could only show that if *w* is possible and derived, it cannot be derived from [*e*]. Notice that this would leave open not just that *w* is derived, but that it is derived *and* means '*e*': the derivational source of *w* might be some phrasal synonym of [*e*].

of 'cowed'. Assuming that *a cow had* is a well-formed meaning, HK's sort of account does not explain why there could not be a primitive lexical item 'cowed' that means *a cow had*. Likewise in the general case: assuming that *e* is a well-formed meaning, the assumption that the derivation of the corresponding expression $\lceil e \rceil$ is blocked does not explain why *e* cannot be the meaning of a primitive lexical item, whether or not $\lceil e \rceil$ is syntactically derivable.

It may be well (in passing) to explicitly distinguish the two kinds of theories that we mentioned in note 2. Presumably $\lceil e \rceil$ is underivable either because the grammar of English lacks the relevant rules or because no grammar of a natural language could have the relevant rules. In the first case, the most that would follow is that, if *w* is a lexical primitive, then it cannot be defined *in English* (i.e., there is no phrase in English that is synonymous with *w*). In the second case, the most that would follow is that, if *w* is a lexical primitive, then it cannot be defined *in any natural language*. In either case, further argument would then be required to show that if the meaning of *w* cannot be defined in natural language, *w* is *ipso facto* an impossible word. (It is not obvious that such arguments would be forthcoming given the empirical plausibility of the claim that the majority of lexically primitive expressions in every natural language are not definable.)

It is important to bear in mind, at this point, what it seems to us that impossible-word arguments are forever forgetting: what is supposed to make a word impossible is that there is something defective about its meaning—that is, what makes it impossible that *w* should mean *e* is that *e* is not a possible meaning. (Patently, if *e* is not a possible meaning, then one does not need an explanation of why no word can mean it.) But how could any conclusion of that sort follow from a demonstration that $\lceil e \rceil$ isn't derivable, from the grammar of a (or any) natural language? Maybe it's derivable in Fortran. We suspect that impossible-word arguments suffer from a use/mention confusion—in particular, confusing 'the expression $\lceil e \rceil$ isn't possible' (which is what instances of the argument actually show when they are successful) with 'the meaning *e* isn't possible' (which is what using impossible-word arguments to defend lexicalization would actually require). There may be sound arguments the conclusion of which has the form '*e* is not a possible meaning' but we don't know what they

would look like; nor have we ever seen one. (Notice that even *round square* is a possible meaning; in fact, it's what 'round square' means in English.)

Impossible Structures

In case 2 a derivation is blocked because the structure it generates is impossible. (The constraint is on the product rather than the process.)

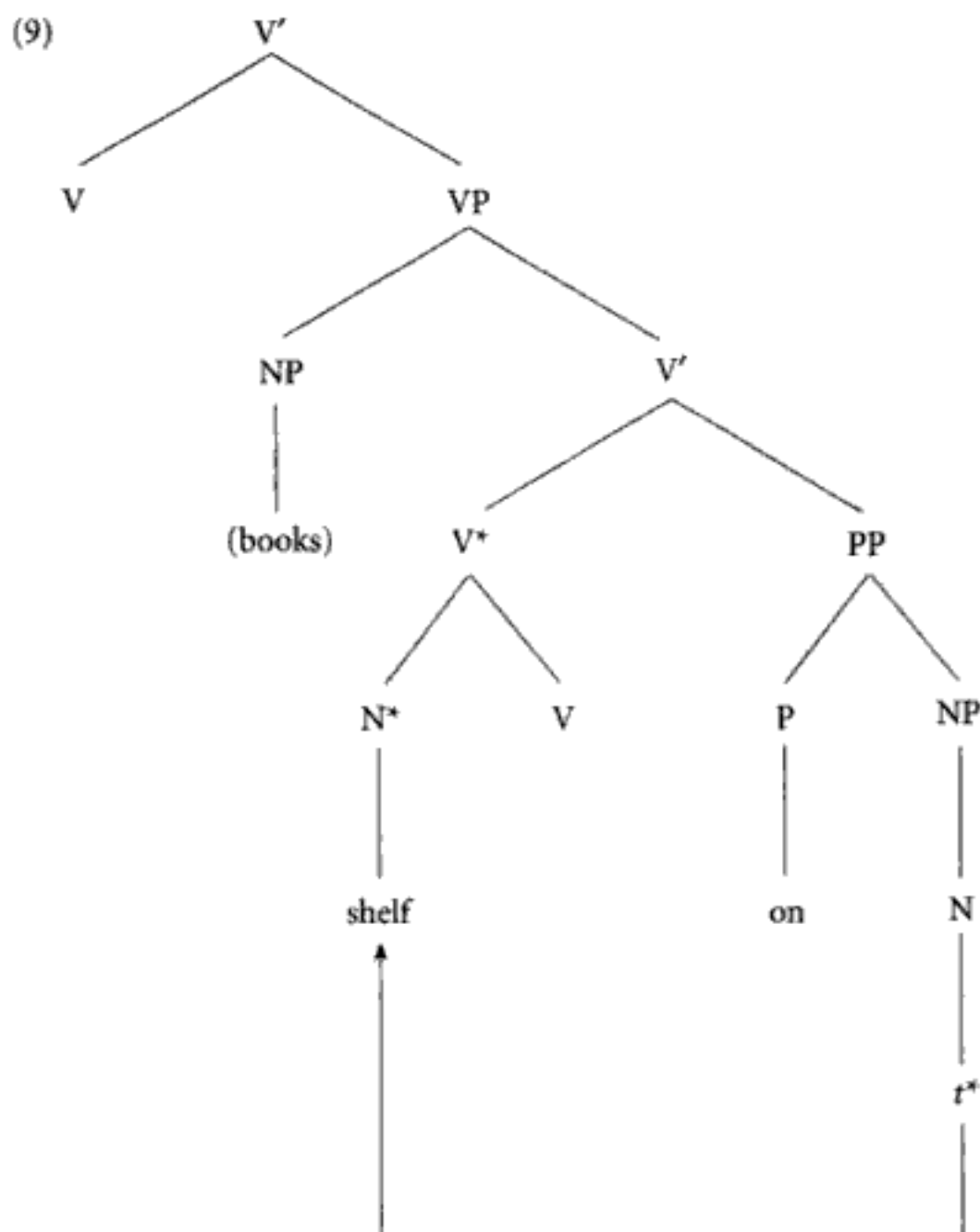
HK do not always assume that what blocks derivations in their examples is a prohibition against moving a certain constituent to (or from) a certain position (e.g., the putative principle that there cannot be a transformation that lowers a subject into a predicate). Instead, sometimes the constraint is on the form of the tree that a derivation *would* produce if it were legal. Here is HK's explanation of why there cannot be a verb 'shelve . . . on' such that (7) is grammatical and means (8).

(7) *He shelved the books on.

(8) *He put the books on a shelf.*⁶

[The hypothetical item] 'shelve (books) on . . .' is derived by incorporation of the noun that heads the complement of the preposition, as shown in (9). The trace of incorporation is thus 'too far' from its antecedent and is therefore not properly governed, violating the empty category principle. Although the trace is coindexed with the verb to which its antecedent is adjoined (as indicated by the asterisk notation), this verb does not govern the trace. The preposition is a 'closer governor', defining PP as the minimal governing domain for the trace By Minimality, therefore, PP is a barrier to government from the more distant verb. (1993: 61)

⁶ Notice that (8) expresses a possible meaning (in fact, it expresses the meaning that 'He put the books on the shelf' actually has). This seems to be a clear case where the question 'What expressions are syntactically derivable?' has been confused with the question 'What lexical items are (semantically) possible?' However, the discussion to follow does not rely on this criticism. (It adds to the confusion that expressions like 'phone up' may well be complex words, that is, lexical items with internal constituent structure. This is not, however, the sort of analysis HK are imagining for 'shelve . . . on' in (7). In particular, the structure HK are considering is one where 'the shelf' is the *object* of the preposition 'on', not one where it is the object of a complex verb 'shelve on'.)



In short, a sentence meaning (8) cannot come from (7) because (7) is illegal; and (7) is illegal because (9) violates a constraint on variable binding that well-formed formulas are *ipso facto* required to comply with. Notice, crucially, that this explanation would account for why (7) is ungrammatical whether or not the putative verb 'shelve . . . on' is supposed to be derived—a *fortiori*, whether or not (7) is supposed to be derived from a sentence that means (8). In particular, the empty category principle would block (7) even if 'shelve . . . on' is primitive and is base-generated. So, in this sort of case, HK's explanation of the 'impossible' form really does explain why it is

impossible. Compare case 1, where the fact that the derivation of (1) is blocked shows only that if 'to cow' is possible, then it must be primitive. So HK's account of case 2 has the virtue that their account of case 1 crucially lacks.

However, it also has the defects of this virtue. Remember that the whole point of making an impossible-word argument is to provide evidence for lexicalization—that is, to provide evidence for inferring that there is a grammatical process that derives surface lexical items from complex, phrasal underlying structures. So, the assumption that lexicalization respects independent constraints on derivations is supposed to explain why intuitively impossible words *are* impossible; and the fact that *if lexicalization obeys independent constraints on derivations, then it explains why impossible words are impossible* was supposed to be the evidence that there is such a process.

But how would this pattern of inference go in the present case? If (7) is ill formed for the reason that HK allege—namely, because there is a mandatory constraint that the structure (9) fails to meet—then, patently, (7) is ill formed whether or not there is a lexicalization transformation. Indeed, on the present assumption, the ill-formedness of (7) depends entirely on features of its geometry; it is, as it were, intrinsic and hence independent of any assumption about how (7) was derived. But if (7) would be ill formed whether or not there is a grammatical process of lexicalization, how could the fact that it *is* ill formed be evidence for such a process? We assume it is not in dispute that *if Q whether or not P, then Q is not evidence for P*.

Conclusion

'Generative semanticists' and 'interpretive semanticists' both hold that some surface lexical items are represented by phrases at (some of) the level(s) of grammatical representation at which semantic interpretations are assigned. The difference is that generative semanticists also think that such surface lexical items are *derived from* the corresponding underlying phrases. The strategy of impossible-word arguments is to provide evidence for the stronger claim by showing that the derivations of certain lexical items would be blocked by independently motivated grammatical constraints and

that, when they are, the corresponding lexical items are intuitively impossible.

We claim, however, that all such arguments are either too weak or too strong to support the thesis that the derivation of words from phrases is a bona fide grammatical process. In one kind of case, it is left open that *underivable* words might not be *impossible* (because it is left open that they might be primitive); in the other kind of case, precisely because the blocked structures are ungrammatical in virtue of their geometries rather than their derivational histories, their being ill formed provides no evidence for or against there being such a process as lexicalization.

But quite aside from the relatively mechanical questions that have primarily concerned us, there is a *prima facie* reason to think that impossible-word arguments must be deeply flawed. In effect, such arguments infer conclusions of the form ‘“e” is not a possible word meaning’ from premises of the form ‘ $\lceil e \rceil$ is not derivable in a certain grammar’. Now it is true that, once upon a time, philosophers dreamed of a ‘logically perfect’ language in which all and only the *meaningful* (*verifiable?*) (*truth-evaluable?*) sentences would be *syntactically* well formed (see, e.g., Wittgenstein in the *Tractatus*). But even philosophers never thought that one could formalize English in such a language; the most they hoped for was to formalize mathematics or physics. In any case, these days everybody thinks that both projects are hopeless. Our point is that the inference from ‘ $\lceil e \rceil$ is underivable’ to ‘“e” is not a possible meaning’ would be sound only on the assumption that $\lceil e \rceil$ is an expression in such a language. So cognitive scientists who endorse impossible-word arguments are thereby committed to a kind of linguistic utopianism that even philosophers are unable to believe.

Barring blatant confusions of use with mention, such facts as that there is no rule of subject raising, or that the empty category principle controls variable binding, say nothing at all about what words are possible. All they tell us is exactly what they seem to—namely, why sentences like (1) and (7) are not grammatical.

Part III

New Pragmatists

Brandom's Burdens: Compositionality and Inferentialism

ROBERT BRANDOM has it in mind to run a 'pragmatist' theory of content(/concept possession/linguistic competence). That is, he wants to reconstruct notions like *saying such and such* or *believing such and such* (and, eventually, notions like *knowing what a word means* and *having a concept*) in terms of a distinctive kind of knowing *how* or being able to *do* something (2000: 17).

'So what?', you might reasonably inquire. 'Isn't practically everybody a pragmatist these days by that criterion?' Well, it's true that lots of people think that 'knowing how' is at the bottom of the heap when questions about content arise. But, not every pragmatist is also what Brandom calls an inferentialist. Most semantic pragmatists hold one or another kind of 'dual aspect' theory about the kind of know-how that content supervenes on. Roughly, they think that grasping content is both a matter of knowing how a word/concept behaves in inferences, and also a matter of knowing how to apply the concept to things in its extension. Having the concept DOG is being able (and disposed) to do a little inferring (e.g., from DOG to ANIMAL); and it's also being able to do a little applying (e.g., of DOG to dogs in conditions that are favorable for dog-spotting). What's striking about Brandom's inferentialist pragmatism is that the applying part more or less drops out in favor of the inferring part.

This, for Brandom, is part and parcel of a wish to emphasize the characteristically human and social character of content. All sorts of things might spot dogs: prelinguistic infants or mere animals for two

cases in point. Even thermostats spot changes in the ambient temperature. But mere differential responding achieves the status of conceptualization only when it becomes inferentially elaborated, thereby providing 'reasons for making other moves in the language game, and as themselves potentially standing in need of reasons that could be provided by making still other moves' (2000: 17). Correspondingly, conceptualization is the typical concern of creatures that give their attitudes linguistic expression, since 'expressing something is *conceptualizing* it . . . in general, addressing it in a form that can serve as and stand in need of reasons, making it *inferentially* significant' (2000: 16).

If there are any concepts for which a pure inferential role semantics seems intuitively appealing, they're the ones that correspond to the logical constants. In effect, Brandom's inferentialist program is to model concept possession at large on what he describes as a Gentzen-style account of logical vocabulary. '[T]he content to which one is committed by using the concept or expression may be represented by the inference one implicitly endorses by such use, the inference, namely, from the circumstances of appropriate employment to the appropriate consequences of such employment . . . [This is] a generalization of a standard way of specifying the inferential roles of logical connectives' (2000: 62). If you find it *prima facie* not plausible that the semantics of, say, 'bird', 'oxidation', 'xylophone', 'zeugma', or 'afternoon tea' is interestingly illuminated by the semantics of 'and', you will not be sanguine that this program will succeed. That these qualms are well founded is the substance of much of what follows.

First, however, a number of geographical remarks.

To begin with, Brandom's pragmatism, though it is reductive, isn't any sort of naturalism. Brandom wants to stand neutral on the question whether it's possible to replace semantic/intentional vocabulary with 'purely physicalistic' terms *salva veritate* (or anything else of interest; 2000: 185–6). An explication of content by reference to 'knowing how' clearly would not decide this issue since the latter is itself up to its ears in intentionality.¹ Rather, what Brandom proposes

¹ Brandom is entirely alert to this. Cf. passages like 'Forming an intention . . . to put a ball through a hoop requires knowing what it is to put a ball through a hoop—what must be *true* for that intention to *succeed*' (2000: 82).

is something like a reduction of the representational to the intentional (or better: to the intentional and the normative): Concepts and the like are individuated by their content, and there's nothing more to having a concept than knowing how to use it. And, whatever exactly knowing how to use a concept may come to, no clauses about representation, satisfaction, truth, or denotation are ineliminably required in the analysis. Notice that this precisely reverses the order of explication that is endorsed by 'representational' theories of mind, according to which it's the vehicles of thought (ideas, or mental representations, or whatever) that have the 'underived' intentionality from which the contents of propositional attitudes and the formulas of public languages are supposed to be inherited.

But if representationalism is the explicit enemy, so implicitly is the Semantic Cartesianism (our term, not Brandom's) of which representational theories of mind are just one species. Anybody is a Semantic Cartesian who holds that *having the concept F is being able to think about F-ness (as such)* and, correspondingly, that *having an expression that means F is being able to express one's thoughts about F-ness (as such)*. According to this sort of theory, the appeal to semantic notions (like 'representation of', and 'thinking about') in accounts of linguistic and conceptual content is ineliminable, short of a naturalistic reduction. The order of exposition, according to all versions of Semantic Cartesianism, is *from the semantic to the intentional*, not the other way around. We're much inclined to think that the Cartesian has the right end of this stick. Brandom's arguments to the contrary haven't shaken us in this conviction.

One more preliminary point. Brandom often writes as though he takes his inferentialism to be continuous with theories of content that he associates with Kant, the young Frege, and Hegel. The idea is, roughly, this: If you're a Cartesian, you're likely to suppose that *thinking about* (e.g., thinking about *redness*) is what's basic to theories of content; *thinking that* (e.g., thinking that *this fire engine is red*) requires, though it does not reduce to, *thinking of* the fire engine and *of* being red. Since, in this order of analysis, *thinking about* is taken to be prior to *thinking that*, referring is likewise taken to be prior to judging. Whereas, if you're an inferentialist, you're likely to think that *judging* is what's primitive, since *judgments* (or their linguistic expressions) are, as it were, the smallest things that can enter into inferences.

Recognizing the priority of judging to other kinds of mental acts is, according to Brandom, what inferentialism has in common with its Kantian, Hegelian, and Fregean precursors.

Now, we don't know about Hegel and, life being short, we aren't expecting to find out. But we're pretty sure that neither Kant nor the young Frege was an inferentialist in anything like Brandom's sense. It's true that both of them thought that judgment plays a fundamental role in the theory of mind; and, correspondingly, that the vehicles of judgment (sentences and thoughts) play a fundamental role in the theory of meaning. Primarily, this was a point against the empiricists who, according to Kant and Frege, had a theory of reference but no theory of predication. In effect, Kant and Frege argued, it is because *thinking about* isn't sufficient for *thinking that* that empiricism is in principle unable to produce an adequate theory of cognition. This is, to be sure, a very important insight; somebody ought to call it to the attention of connectionists. But it doesn't make Kant or Frege an inferentialist because, simply, 'thoughts are the units of inference' is entirely compatible with 'concepts are the constituents of thoughts'. Likewise for language. Strawson, for example, held that referring is a speech act and that it's prior to asserting in the order of analysis. Why isn't this a Kantian view? What Kant and Frege did for the theory of mental representation was to *add predication*, not to *subtract reference*.

Here's another way to put this point: Kant might well have, and Frege would have, agreed that identity of inferential role is sufficient for identity of content. (Frege does so explicitly in section 3 of his *Begriffsschrift*.) But that wouldn't, of course, imply that inferential role is primitive in semantic explanations. In fact, as we read them, Kant and Frege *couldn't* coherently have been inferentialists. This is because (unlike Brandom; see below) they had an answer to the question '*which* inferences does content supervene on; which inferences *constitute* a concept's inferential role?' Their answer was that content supervenes on *analytic* inferences; to know what 'dog' means is, *inter alia*, to know what inferences follow from '... is a dog' in virtue of the *meaning* of the predicate. (Likewise, *mutatis mutandis*, for having the concept DOG.) So, as we read them, though Kant and Frege held that the content of a concept supervenes on its inferential role, they also held that the inferential role of a concept supervenes on its

analytical connections. Since analyticity is *truth* in virtue of *meaning*, this makes the semantic properties of a concept prior to its inferential role. Kant and Frege were good Semantic Cartesians after all. (To say nothing of their being good Semantic Platonists.)²

So much for the scholarship; now to the dialectics. There are two key questions that theories of content must answer, and which have generally been supposed to be hard for inferential role accounts of meaning. These are: *Which inferences are meaning-constitutive?* and *How do you explain the compositionality of meaning?* A semantics that fails to provide these questions with adequate answers is, we think, *ipso facto*, not tenable. Let's, then, see what Brandom's inferentialism has on offer.

Which Inferences?

Brandom's semantics says that the content of an expression (/thought) is(/supervenes on/is 'determined' by) (2000: 29, 47) its inferential role. This is, however, a darker doctrine than it may at first appear. Thus, *prima facie*, inferentialists mustn't hold that *identity* of inferential role is a necessary condition for identity of content. For example, whereas it looks like the belief that P and the hope that P' have the same content (viz., P), their causal-cum-inferential roles are surely radically distinct.³ But nor may inferentialists

² It would be much in the Kantian spirit to suppose that you can't *think* of something without having recourse to some category or other; this might be a generalization of the idea that there is 'no seeing without seeing as'. But that would be entirely compatible with the metaphysical thesis that what you're *thinking that* depends on what you're *thinking of* in the sense that the vehicles of reference are the constituents of the vehicles of inference. What's prior to what depends a lot on what kind of priority you have in mind.

³ There are other reasons that are arguably more pressing. As Brandom rightly remarks, 'Inferentialism of any sort is committed to a certain kind of semantic *holism* as opposed to the *atomism* that often goes hand in hand with commitment to a representationalist order of semantic explanation. . . . Such holistic conceptual role approaches to semantics potentially face problems concerning both the *stability* of conceptual contents under change of belief and commitment to the propriety of various inferences, and the possibility of *communication* between individuals who endorse different claims and inferences' (2000: 29). We have elsewhere emphasized the dire consequences that semantic holism has for intentional explanation of any sort (see Fodor and Lepore 1992) and we don't take back a word of it. But we won't

hold that the content of an expression supervenes on its role in *purely logical inferences* (i.e., on inferences that turn just on the meanings of the logical constants); for these do not, in general, distinguish between formulas of the same logico-syntactic type (for example, between 'that's a cat' and 'that's a mat'). The typical way of running a conceptual role semantics is therefore to distinguish a class of *analytic* inferences, and say that they, and only they, are constitutive of the concepts that they contain. Thus, if (but only if) it's analytic that cats are animals, then the inference CAT $\rightarrow$ ANIMAL is constitutive of one or other (or both?) of the concepts.

But this conventional way of solving the 'which inferences?' problem isn't, of course, available to Brandom. As we remarked a few paragraphs back, analyticity is supposed to be *truth in virtue of meaning*. For the content of a concept to supervene on its role in *analytic* inferences is thus for the inferential role of that concept to supervene on its semantics. This is, to be sure, the order of dependence that a Cartesian assumes if he recognizes a notion of analyticity at all. Barring familiar Quinean scruples, it seems to us clearly the right way to proceed. Inferentialism, however, is prohibited from taking this route, since the inferentialist's point in a nutshell is that meaning is a construct from inferential role, not the other way around.

So, then, what is Brandom's answer to 'which inferences?' Well, actually, he doesn't say. The idea, as far as we can make out, is that conceptual mastery 'is not an all-or-none affair; the metallurgist understands the concept *tellurium* better than I do, for training has made her master of the inferential intricacies of its employment in a way that I can only crudely approximate' (2000: 64). Or, as he puts it elsewhere, 'Concepts . . . must come in packages (though it does not yet follow that they must come in just one great big one)' (2000: 16).

dwelling on that now, because Brandom has little to say about such concerns except that they 'are rendered much less urgent . . . if one thinks of concepts as *norms* determining the *correctness* of various moves' (2000: 29). Actually, we don't understand that very well. Suppose you think that DOG implies ANIMAL, and we think that it doesn't. Surely, a difference in norms is *thereby* implied, since whereas you think it's *permissible* to argue from DOG to ANIMAL, we think that such inferences *ought not* to be countenanced. (You might say: 'Well, that kind of difference of norms doesn't count'; but that only replaces the 'which inferences?' problem with the 'which norms?' problem, for a total gain of no yardage.) We won't press the holism worries in what follows; but we do urge you to keep them in mind in evaluating the inferentialist program.

Maybe all of this is so,⁴ but of course it begs such questions as what it is, in the situation imagined, that determines whether *the metallurgist's concept of tellurium is the very same one that I am struggling to grasp*. Since this seems to be just the 'which inferences?' question all over again, the example doesn't help with the problem at hand.⁵

In any event, as far as we can tell, Brandom is committed to holding that at least some of the *material* inferences in which an expression is implicated are constitutive of its content. This emerges in passages like the following:

The kind of inference whose correctnesses [*sic*] determine the conceptual contents of its premises and conclusions may be called, following Sellars, *material* inferences. As examples, consider the inference from 'Pittsburgh is to the west of Princeton' to 'Princeton is to the east of Pittsburgh,' and that from 'Lightning is seen now' to 'Thunder will be heard soon.' It is the contents of the concepts *west* and *east* that make the first a good inference, and the contents of the concepts *lightning* and *thunder* . . . that make the second appropriate. Endorsing these inferences is part of grasping or mastering those concepts. (2000: 52)

This commitment doesn't, of course, answer the 'which inferences?' question, because we still don't know which *material* inferences are crucial. But notice, in any case, the strain that invoking material inferences in this context puts on the idea that inferential role can reconstruct notions like concept possession.

It may be a matter of linguistic rule, or social convention, or whatever, that determines whether dogs have to be animals. But only God can determine that lightning is reliably followed by thunder. Whether thunder reliably follows lightning has *nothing to do* with which inferences I, or my society, or even the experts I defer to, take to be (in Brandom's term) 'appropriate'.⁶ Thunder follows lightning not

⁴ But also maybe it's not. We would have thought that the relevant difference between us and the metallurgist is not that she has a better grasp on the concept *TELLURIUM* than we do, but that she *knows more about tellurium* than we do (come to think of it, who doesn't?). Semantic assent is a fine thing in its way, but it does not, as one says in Australia, cure wooden legs.

⁵ Perhaps, in the case of 'tellurium' and the like, it has something to do with patterns of deference; but natural kind terms aren't the general case; and, anyhow, a pattern of deference isn't an inferential role.

⁶ Brandom says 'appropriate' where Semantic Cartesians would say 'true'. Embarrassing questions like 'are appropriate assertions *ipso facto* true?' and 'are false assertions *ipso facto* inappropriate?' never get raised; perhaps as a matter of principle.

because that's the way we play the language game but because of the laws of meteorology. Surely, what inferences are *materially* good is a matter of how the world is; *and, surely, we are arbitrarily ignorant of how the world is*. All this being so, the more that the material inferences that a concept is involved in are supposed to constitute its inferential role—and hence, its content—the less *grasping* a concept can be identified with mastery of its inferential role. If, at the limit, it's supposed that grasping a concept requires mastery of *all* the corresponding material inferences, then only God has any concepts.⁷

The score so far: Like everybody else who thinks that content comes from inferential role, Brandom needs, but doesn't have, a story about which inferences are the ones that content comes from. What Brandom adds to the general perplexity is the idea that these include material inferences *inter alia*. That makes it hard to see how Brandom can acknowledge the truism that lots of people share lots of concepts. But, still worse, it precludes him from distinguishing what the world contributes to the reliability of inferences from what the (putative) rules of language contribute. We don't, ourselves, think that languages are much like games. But, even if we did, we wouldn't want 'thunder after lightning' to be a linguistic rule. 'Analytic ethics' and 'analytic epistemology' and the like were quite bad enough; the prospect of an analytic meteorology is really more than we can bear.

Compositionality

Here are some propositions that practically everybody agrees on, including Brandom as we read him (see 2000, ch. 4).

- (1) Human linguistic and cognitive capacities are productive; and the explanation of this fact turns crucially on their being compositional.⁸

⁷ Though he doesn't say so, it may be Brandom's view that, insofar as material inferences are supposed to be content-constitutive, the crucial semantic relation among concepts is *similarity* rather than *identity* of content. Many inferential role semanticists who are wary of appeals to analyticity have suggested this way of dealing with holism problems. But here the devil is in the details and, so far as we know, there aren't any. (See Ch. 9 in this volume.)

⁸ Likewise, minds and natural language are invariably systematic. Brandom doesn't

- (2) As far as anyone knows, compositionality means that both the syntax and the semantics of complex linguistic expressions like sentences (and, *mutatis mutandis*, complex mental representations like the ones that express thoughts) are determined entirely by the syntax and semantics of their primitive constituents.
- (3) Compositionality is therefore 'not negotiable'; a theory of content that can't accommodate a notion of compositionality adequate to the explanation of productivity is *ipso facto* unacceptable.

We take these considerations very seriously; and we have argued elsewhere that they preclude not just inferentialism, but *any* theory according to which inferential role is content-constitutive. (See Fodor and Lepore 1992.) This is because, according to us, inferential role is itself noncompositional: in general, the inferential role of a sentence/thought is not determined by the inferential roles of its constituents. Brandom doesn't consider this objection at all. Since we have already made it frequently, loudly, and in many places, we propose to waive it here and consider just those aspects of the compositionality problem that Brandom does discuss.

To begin with, we assume that it's implicit in (2) that sentences are constructs out of subsentential expressions and, *mutatis mutandis*, that thoughts are constructs out of constituent concepts. This is, in principle, something that a Cartesian is entirely comfortable taking for granted. No doubt Cartesians, like everybody else, are hard put actually to provide an account of how compositionality works. It is, indeed, possible to view the last fifty years of research on the syntax and 'formal semantics' of natural languages as propaedeutic to such an account. Our point is just that, since Cartesian Semantics takes *reference* to be prior to *truth*, and *thinking about* to be prior to *thinking that* (see above), there is no reason why it shouldn't likewise take the semantics of sentences and thoughts to be derivative from the semantics of words and concepts.

The inferentialist's situation is, however very much different.

discuss systematicity, but the relevant considerations run parallel to the ones about productivity. See Fodor and Pylyshyn (1988).

Remember that he is committed to the metaphysical and explanatory priority of *judgments* to other mental acts. Since the typical linguistic vehicle for expressing a judgment is a sentence, it is thus sentences, rather than their constituents, over which content properties are defined in the first instance. So, before he can even start on the compositionality problem, Brandom has to make clear how the semantics of subsentential expressions might be determined by the inferential roles of their sentential hosts. 'The linguistic community determines the correct use of some sentences, and thereby of the words they involve, and so determines the correct use of the rest of the sentences that can be expressed [*sic*; "constructed"?] by using those words' (2000: 128).

In fact, quite understandably, Brandom discusses only one aspect of this problem. He wants to show that, starting with the inferential role of sentences, and assuming no subsentential semantics or syntax, he can isolate and appropriately characterize the class of singular terms, either in linguistic or in mental representations. That would amount to less than an inferentialist account of compositionality, but we agree with Brandom that achieving it would greatly strengthen the case for inferentialism. So, it's very important, from our point of view, whether the project succeeds. Well, it doesn't; as we will now see in some detail.

In effect, Brandom has a two-step definition of 'singular term' and, as far we can tell, neither step works. The first move is to use the Fregean notion of substitution to define a grammatical class of subsentential expressions that is, according to Brandom, guaranteed to be a (possibly improper) superset of the singular terms, thereby deriving their syntax from the underlying notion 'well-formedness-preserving inference'. Given the kind of grammar that Brandom has in mind, the notion of membership in a syntactic category is reasonably clear: 'Two subsentential expressions belong to the same syntactic or grammatical category just in case no well-formed sentence . . . in which the one occurs can be turned into something that is not a sentence merely by substituting the other for it' (2000: 130). That is, the grammatical categories are conceived of as a set of intersubstitutable segments of sentences (where a sentence is something that has a characteristic sort of speech act potential; in partic-

ular, a characteristic inferential potential in the case of declarative sentences).⁹

We think we follow the argument so far, in that we're prepared to assume that a grammatical category is essentially a set of expressions that can substitute for one another in a sentence, *salva congruitate* (that is, without destructive effect upon the grammar of a sentence. Actually, we will take back this concession later on; but one thing at a time). Since, however, this is a perfectly *general* notion of grammatical category, we don't see how it helps with the problem of isolating the *particular* grammatical category to which singular terms belong. After all, substitution for an adjective, or an adverb, or a determiner, or a preposition . . . will also generate a grammatical class. For example, you can think of the (English) adjective as the grammatical category generated by substituting for 'red' in 'John ate a red apple'. But, of course, the singular terms don't belong to that class. Rather, the category of singular terms is what you get by substituting in a sentence *for a singular term* (e.g., for 'John' in 'John ate a red apple'). On pain of circularity, however, that is of no use to Brandom, whose aim is a procedure for *introducing* 'singular term' into a theory where the only primitives are 'sentence' and 'well-formedness-preserving substitution'.

In effect, our point is just the obvious one that recursive definitions need starting axioms, and starting axioms need justifications. Suppose you think that NPs are the expressions that substitute in the frame '___ PRED' to produce sentences. You thereby assume not just the notion *sentence*, but also the notion *predicate*. To be sure, you can avoid this by just stipulating that some expression *e* is a predicate; and you can then proceed, in the usual style of recursive definition, to identify the NPs as the set of expressions that substitute *salva* sentencehood in the frame '___ *e*', and the predicates as the set of expressions that substitute *salva* sentencehood in the frame 'NP ___'. This defines 'PRED' and 'NP' relative to 'S' *together with the assumption* that *e* is a predicate; so, if you were to pick as ^[e] some expression that is *not* a predicate ('of', as it might be), you will get misdefinitions of 'NP'

⁹ Notice that, strictly speaking, Brandom allows himself two fundamental notions in the reconstruction of 'singular term'; viz., *inferential potential* and *substitution that preserves well-formedness*. Presumably, neither of these reconstructs the other.

and 'PRED'. (For example, the NPs are not identifiable with the expressions that substitute *salva* sentencehood for 'He is the man I showed you a picture' in 'He is the man I showed you a picture of'.) This is a point that Chomsky made with considerable effect against taxonomic linguists some four decades ago; they too had thought that appeals to substitution in sentences would ground a 'discovery procedure' in terms of which 'NP', 'PRED', and the like could be defined. As far as we know, however, nobody in linguistics thinks that any more.

So much on the assumption, implicit in Brandom's appeal to the substitution test, that grammatical classes form a hierarchy of sets of subsentential expressions. In fact, however, they don't; ask any linguist (or see Chomsky 1965). Consider, for example, the class of nouns (to which, presumably, singular terms *ipso facto* belong). You might suppose that they are a proper subset of the NPs; e.g., the ones that substitute for 'John' in 'John ate his own lunch'. But that doesn't work because *'Mary ate his own lunch', and 'Mary' is surely a noun. You could try identifying the nouns with the NPs that substitute for 'John' in 'John ate his or her own lunch' since 'Mary' is among those. But notice *'They ate his or her own lunch'. In fact, there is no way to represent the distinctions ('masculine noun'/'feminine noun') and ('singular noun'/'plural noun') as a hierarchy; *all* the arrangements of these categories are possible. The moral usually drawn is that the singular nouns aren't a *subset* of the nouns; rather, they are the nouns that bear the feature '+singular'. Patently, the taxonomies that feature assignments can generate are a superset of the hierarchical taxonomies. So, it's a serious mistake to rest a theory of grammatical categories on a notion of well-formedness that permits only taxonomies of the latter kind.¹⁰

Pace Brandom, it's not possible to define 'Subsentential English wff' in terms of 'sentence' and 'substitution'. In effect, you have to go the other way around; start with the notion of a subsentential English wff, and define 'English sentence' in terms of it. Just as Cartesian accounts of compositionality have always thought, the units of syn-

¹⁰ Of course, there needn't be anything wrong with applying a hierarchical notion of grammatical category to some or other *regimented* language. It may be that Brandom is implicitly assuming that the syntax of English works like the syntax of typical formalizations of quantificational logic. But it doesn't.

tactic analysis are subsentential; this is so even though it's whole sentences that express the units of *thought* (and inference).

Suppose, however, that we just give the inferentialist a syntax of English that specifies the well-formed subsentential expressions. Can we *then* proceed to identify the *semantically* singular terms without also presupposing notions like reference, denotation, and the like? That would be no small trick even if, having performed it, the inferentialist still is in debt for a notion of syntax that his theory of compositionality is, *pro tem*, unable to reconstruct.

A Cartesian will bet that he can't. That's because Cartesians suppose that subsentential expressions are the units, not just of syntactic analysis, but of semantic analysis as well. In particular, Cartesians think that whether an expression is a singular term depends on the symbol-world relations it enters into, not (or not just) on the inferential roles of the symbols that contain it as a constituent. Well, we don't have a proof that you can't identify singular terms in a way that would satisfy inferentialist scruples, but we're pretty clear that Brandom's idea for doing so won't work.

Brandom's proposal is to use a substitution test: singular terms are all and only the ones for which the schema S_1 is valid:¹¹

$$S_1: (Fx \rightarrow Fy) \text{ iff } (Fy \rightarrow Fx).$$

Thus, for Brandom, it is part and parcel of 'Mark Twain' and 'Samuel Clemens' being singular terms that, if you are prepared to infer from 'Mark Twain is the author of *Tom Sawyer*' to 'Samuel Clemens is the author of *Tom Sawyer*', then you should likewise be prepared to infer from 'Samuel Clemens is the author of *Tom Sawyer*' to 'Mark Twain is the author of *Tom Sawyer*'. Contrast 'This is red, therefore this is colored', an inference that is not, of course, valid if you substitute 'red' for 'colored' and vice versa. (For purposes of the discussion, we assume that definite descriptions are singular terms; many of the examples that Brandom gives to show how singular terms work are in fact definite descriptions. We're not at all clear that this is the right treatment of definite descriptions; perhaps they are quantifiers (Russell 1905). But we won't argue the point here.)

¹¹ Thus, as Brandom remarks, both the syntax and the semantics of singular terms are determined by their behavior in inferences that involve substitution; the difference is whether it's truth or well-formedness that the substitutions preserve.

It is, to put it mildly, not obvious why anyone should expect that the singular terms are identifiable as the ones for which S_1 is valid. It may be that Brandom is assuming that, whereas inferences that substitute singular terms depend fundamentally on *identity*, inferences that turn on predicates depend fundamentally on notions like set inclusion. The point would be that S_2 is valid (barring intensional contexts), but S_3 is not.

S_2 : $a = b$ iff $(Fa \text{ iff } Fb)$

S_3 : $a \in b$ iff $(Fa \text{ iff } Fb)$.

Whether or not this diagnosis is correct, however, it seems clear that there are counterexamples to Brandom's account. Thus, there are inferences involving singular terms that are valid in one direction but not in the other because they turn on part-whole relations. Consider 'John painted his whole arm red' $\rightarrow$ 'John painted his wrist red', which is only valid in that direction. To be sure, this example assumes that 'his whole arm' and 'his wrist' are singular terms, which is in the spirit of Brandom's assumption that definite descriptions are. But suppose that's wrong. Still, 'this' and 'that' are singular terms if anything is a singular term; but 'This is entirely red' $\rightarrow$ 'That is entirely red' can be valid even if 'That is entirely red' $\rightarrow$ 'This is entirely red' is not, as in the case where 'this' refers to my arm and 'that' refers to my wrist. Likewise, with paradigm cases of names. 'Father was at Magdalen' implies 'Father was at Oxford' but not vice versa.

Indeed, as far as we can see, there could be all sorts of inferences which (as a Cartesian would put it) involve relations among individuals but aren't biconditional. Here's a quickie that isn't about parts and wholes: a little glass in a big glass (right side up). If the big glass is filled with water, then the little glass is too; but not the other way around.

And so forth.

Having argued, to his own satisfaction, that the singular terms can be identified by reference to his substitution test, Brandom then goes on to argue that conditional reasoning and reasoning from negations both presuppose expressions that are singular terms by this criterion. If we're right about his test for singularity not working, however, then the most these arguments can show is that if a language can express conditionality and/or negation, then it must contain expressions for

which S_1 is valid. This result is interesting if it's right, but clearly it can't underwrite the sorts of metaphysical conclusions that Brandom takes himself to have established; e.g., that rich languages must acknowledge a notion of object; more generally that 'The limits of [my] language . . . means the limits of the world' (2000: 155).

Summary So Far

As far as we can see, inferentialism can't answer the question 'Which inferences are concept constitutive?', and it likewise can't answer the question 'How does compositionality work?' It may be, of course, that *no* theory of meaning can answer either of those questions, in which case we are all in deep trouble. But our point is that they are intractable for inferentialists *qua* inferentialists. In the first case, that's because even if content supervenes on inferential role, it doesn't follow that inferential role is metaphysically prior to content. For that matter, even if you assume that its inferential role is an *essential property* of a concept, it patently doesn't follow that concepts are, metaphysically speaking, constructions out of their inferential roles. (Presumably, it's necessary that sisters are siblings. It doesn't follow either that SISTER is a construct out of SIBLING or vice versa.) Likewise, even if 'thinking that' is metaphysically prior to 'thinking about', in the sense that you can't do any thinking about unless you can also do some thinking that, it's perfectly possible that the semantics of sentence-length thoughts should be constructed from the semantics of their conceptual constituents.

It strikes us as plausible that a semantics that can't do 'which inferences?' and can't do compositionality is likely to get a lot else wrong too. We will conclude by discussing, relatively briefly, three subsidiary issues that Brandom raises: the *de re/de dicto* distinction, the status of reliabilism in epistemology, and the role of normativity in the analysis of content.

De re/De dicto

Most people think that there is an important distinction between the belief-attributing sentences that are true *de dicto* and the ones that

are only true *de re*. This distinction is supposed to show in such ambiguities as: 'John believes Ortcutt is a spy'. Like Brandom, we think it's important that a semantic theory be responsive to these intuitions.

They are, however, *prima facie* hard for an inferentialist to accommodate. That's because it's so natural to treat the *de re* reading of a belief ascription as one that remains *true* under substitution of *coreferring expressions*; and, *prima facie*, inferentialists have no access to a robust notion either of truth or of reference. Brandom has, on offer, an alternative construal of the distinction; one he says that he shares with all and only Dan Dennett. In fact, we're pretty sure it's wrong.

The basic idea is that

the distinction between *de dicto* and *de re* should be understood to distinguish not two kinds of belief or belief contents, but two kinds of ascription . . . My suggestion is that the expressive function of *de re* ascriptions of propositional attitudes is to make explicit which aspects of what is said express commitments that are being attributed and which express commitments that are undertaken. The part of the content specification that appears within the *de dicto* 'that' clause is limited to what, according to the ascriber, the one to whom the commitment is ascribed would . . . acknowledge. . . . The part of the content specification that appears within the scope of the *de re* 'of' includes what . . . is acknowledged [by the ascriber] as an expression of what the target of the ascription is committed to. (2000: 176–7)

For all we know, this may be a correct account of some of the pragmatics of *de re* and *de dicto* ascriptions; perhaps that is all that Brandom means to claim for it. What seems clear, however, is that you can't explicate the semantic notion of a *de re* reading by reference to the part of the ascribed belief for which the attributer is (or even should be) prepared to take responsibility. Consider John, who says that 'Bill believes the present king of France is bored'. Question: Can the *de re* reading be the one that attributes 'the present king of France' to John? Answer: Clearly not; they aren't even equivalent. We assume, after all, that *de re* ascriptions do remain true under substitution of coreferring expressions, whether or not it's appropriate for semantics to identify them by adverting to the fact that they do; surely, to give this up would be simply to change the topic.¹² So, then,

¹² A propos: it would also burden Brandom with eccentric accounts of *object* and *objectivity*, since he wants to define these by reference to the *de re* readings of propositional attitude attributions.

the result of substituting in the *de re* reading is true whatever true description of the current king of France you put in place of 'the present king of France'. In particular, it remains true whether or not the substituted description is one that the ascriber is prepared to acknowledge.

'Right,' you reply; 'but that's only because you've misunderstood the view. What's claimed is that the *de re* reading holds for and only for those descriptions that the ascriber *should* accept. *De re* is a normative notion too.' This presumably means that the *de re* ascriber is making some sort of mistake—misusing the language—whenever he rejects an ascription that differs from his own only in the substitution of a coreferring expression. Well, if that's the norm, then only God is without sin. Excepting God, nobody can (hence, nobody is required to) identify the class of descriptions referentially equivalent to his own, and we take it that there can't be a norm that nobody but God could comply with. This should be particularly embarrassing for semanticists who think that *material* inferences are constitutive of inferential roles, since it apparently implies that (barring accidents) nobody could use a singular term correctly unless he knew *all* the true identity statements in which the term is implicated.¹³

Reliabilism

Suppose 'John's belief that *a* is *F* is rational' could be analyzed as something like 'John's belief that *a* is *F* was arrived at by a reliable belief-producing procedure'. This would be bad for Brandom, because it would mean that even if mind-world relations don't enter into the analysis of semantic notions like *content*, they do enter into the analysis of epistemological notions like *rational belief*.

In fact, though Brandom thinks that there's a lot that's right about

¹³ Unsurprisingly, the same sort of point also holds about how quantifying in works with respect to *de re* belief ascriptions. Suppose you believe that Vulcan exists. Then, if Brandom's story is right, I could presumably express something true with the form of words 'Bill believes of Vulcan that he exists', where this means that I take responsibility for ascribing to Bill a *de re* belief about Vulcan.

reliabilism, he also thinks that it can't be literally true. He has several reasons for holding this, but the crucial one is that reliability is *context-sensitive*; there is no general context-free notion of reliability in terms of which a corresponding context-free notion of rationality can be reconstructed.

. . . the reliability of the belief-forming mechanism . . . varies depending on how we describe the mechanism and the believer. [Suppose pseudo-barns are frequent in Barn Façade County, but that elsewhere they are rare. Then] described as apparently perceiving this barn, [the observer] is reliable and knows that there is a barn in front of him. Described as an apparent barn-perceiver in this county, he is not reliable and does not know there is a barn in front of him . . . (2000: 116)

And so forth.

But this can't be a serious objection to reliabilism since, though the reliability of perception is indeed context-relative, so too is the reliability of every other kind of instrument that's deployed in rational belief fixation. It is, *ceteris paribus*, rational to believe that the room is hot if the thermometer reads 105 degrees. That's true even though the reliability of thermometers for measuring temperature is context-relative; they don't work on the sun, and they don't work near absolute zero. The moral of the context-relativity of the reliability of thermometers is *not*, however, that using them to acquire a belief about the temperature is insufficient to rationalize the belief. Rather, it's that, insofar as it's an epistemically interesting notion, reliability is really a three-place relation; viz., between the context of measurement, the parameter that's measured, and the instrument that's employed to measure it. This being so, it's reasonable to hold that a procedure for belief acquisition (perceptual or otherwise) might be sufficient to justify a belief in a certain circumstance even though it would not have been sufficient to justify that same belief had the circumstances been otherwise. We see no reason at all why this concession should make a reliabilist unhappy. Compare: the fact that *the very thought that is true in some circumstances may be false in others*; that this is so does not impugn the notion of truth. In particular, it doesn't impugn the idea that truth is a condition on a belief counting as knowledge.

Normativity

This will be a very short section. Apparently, Brandom thinks that the essential insight of inferentialism is that content is a *normative* notion. For content is to be explained in terms of knowing-how; and knowing-how invokes rules, procedures, and the like; and rules, procedures, and the like are the kinds of things that can be applied rightly or wrongly. (See 2000, *circa* 162.) This is fine with us. But we're puzzled by what appears to be Brandom's assumption that you can't make a comparable point if you approach the notion of content from such starting points as truth, truth condition, and the like. As we remarked above, truth is itself a normative notion; it's what you *ought* to believe, all else equal. And, for better or worse, it's part and parcel of contemporary Cartesianism to suppose that truth conditions are all you need to reconstruct notions like representation (see, e.g., Davidson 1984). (Representation is also a normative notion, since it's for sure that some of your representations are *misrepresentations* (unless you are God)).

So, what *is* all this stuff about normativity?

Conclusion

Brandom says of some of his main conclusions that they are 'odd and marvelous'.¹⁴ And indeed they would be if only they were true. As things stand, the relevant points against 'inferentialism' are banal and obvious: You can't rely on the notion of inference to abstract the world from the theory of meaning. This is because you can't abstract

¹⁴ For example: 'we can say that two assertible contents are *incompatible* in case *commitment* to one precludes *entitlement* to the other' (Brandom 2000: 194). Whereas, we would have sworn that two statements are incompatible just in case it's necessary that if one is true, the other is false. You can no more reduce (in)compatibility to an epistemic notion like entitlement than you can reduce truth to an epistemic notion like assertibility; not, anyhow, if it's part of the story that everybody is obliged to know the entitlement conditions of the assertions they can make. To suppose that you can reduce truth to assertibility *by* reducing (in)compatibility to entitlement strikes us as pretty close to begging the question. Compare Brandom (2000, ch. 6).

the notion of truth from the notion of inference; truth is what good inferences preserve, and truth is a symbol–world relation.¹⁵ Likewise, if the language that you talk/think in happens to be productive, then you can't subtract the notion of reference from the notion of truth, since productivity implies the priority of subsentential semantics to sentential semantics. And, finally, you can't subtract the notion of an object from the notion of reference because, in the paradigm cases, objects are what subsentential expressions refer to.

Will anybody who finds any of this in the least surprising please raise his/her hand?¹⁶

¹⁵ N.B., *truth*, not commitment, or even rational commitment. You can't be committed to all and only the good inferences unless you know exactly which inferences the good ones are; which you don't (and neither does 'society'; *pace* Hegel). That is, of course, the classical reason for doubting that *any* form of pragmatism can be made to run.

¹⁶ We would like to thank Robert Brandom, Carlo Penco, Alex Oliver, and Rob Stainton for comments on earlier drafts of this paper.

Paul Churchland and State Space Semantics

A NUMBER of philosophers believe that the notion of identity of meaning should be replaced by some graded notion of similarity of meaning. With a graded notion of similarity of meaning, intentional generalizations would be viewed as subsuming individuals in virtue of the similarity of their mental states; translation and paraphrase would be viewed as preserving not identity but similarity of meaning, and so forth. The prospects, however, for constructing a 'robust' notion of similarity of meaning—one adequate to the purposes of semantics and cognitive science—are, we believe, remote. This prognosis isn't widely shared, perhaps because friends of semantic similarity have generally been careful not to say what they take semantic similarity to consist in. Paul Churchland, however, has proposed a sketch of the sort of similarity theory that is required, a notion of mental (or, anyhow, neural) representation that 'embodies . . . metrical relations . . . and thus embodies the representation of similarity relations between distinct items thus represented' (1989b: 102).

Since Churchland's attitude towards the intentional/semantic generally tends to be eliminativist, it's unclear just what properties of contentful states his 'state space' representations are supposed to preserve. Suffice it that, when he's in a 'highly speculative' mode, he contemplates, for example, the possibility of 'a way of representing "anglophone linguistic hyperspace" so that all grammatical sentences turn out to reside on a proprietary hypersurface within that hyperspace, with the logical relations between them reflected as spatial relations of some kind . . . [This would hold out] the possibility of an alternative to, or potential reduction of, the familiar Chomskyan

picture' (1989b: 109). This entails that state spaces can represent grammars and such. Like much else that he says, it certainly sounds as though Churchland has in mind a kind of representation that specifies the contents of neural states, in which case he is into intentionality up to his neck. In any event, we propose to read him that way and ask how much of the intuitive notion of content similarity state space representation allows us to reconstruct. In what follows, we hope to convince you that, for all that's on offer so far, the problem of semantic similarity still appears to be intractable.

State Space Representation

Here's one of Churchland's formulations: 'The basic idea . . . is that the brain represents various aspects of reality by a position in a suitable state space; and the brain performs computations on such representations by means of general coordinate transformations from one state space to another' (1989b: 78–9).

For our present purpose, which is semantics rather than the theory of mental processes, only Churchland's account of neural representation need concern us. We commence by trying to make clear how Churchland's state space proposal connects with the more familiar 'network' picture of semantics. Churchland's state space proposal is, we suggest, profitably viewed as an attempt to generalize 'network' semantics and to free it from its specifically empiricist assumptions.

Suppose we start with a roughly Quinean account of the structure of languages (theories/belief systems), according to which there are two sorts of ways in which the (nonlogical) symbols belonging to an object language get semantically interpreted. The semantics of the 'observation vocabulary' is fixed by conditioning (or other causal) relations between its expressions and observable properties of the distal or proximal environment. The semantics of the rest of the vocabulary of the object language is fixed by a network of inferential or (in case the semantics is intended to be naturalistic) causal/associative relations to one another and to the observation terms. In effect, the semantic theory of a language thus represents its vocabulary as nodes in a network, the paths of which correspond to semantically relevant relations among the vocabulary items. Observation

terms are at the 'periphery' of the network; nonobservational vocabulary is further in. We take it that this geography is familiar.

Recall how the problem of content identity arises on this 'network' picture. If the paths to a node are collectively constitutive of the identity of the node (as presumably they will be if, as Quine and Churchland both hold, no analytic/synthetic (a/s) distinction is assumed), then only identical networks can token nodes of the same type. Identity of networks is thus a sufficient condition for identity of content, but this sufficient condition isn't robust; it will never be satisfied in practice. In short, a network semantics offers no robust account of content identity if it is denied access to an a/s distinction. But maybe a network semantics can nevertheless be made to offer a robust notion of content similarity? The present proposal is to make content similarity do the work that content identity did in semantic theories that endorsed the a/s distinction.

How, then, might a robust metric of content similarity be constructed; one that is defined for nodes belonging to networks that are (perhaps arbitrarily) different from each other? An immediate problem is this: according to the usual understanding, the only fixed points in a network are the nodes that correspond to observation vocabulary. It's only these peripheral nodes that can be identified without specifying the rest of the network that contains them. (We're supposing that we know what it is for two arbitrarily different theories to both have a node that expresses an observable property like, e.g., red; it's for both to have nodes that are appropriately connected—for example, conditioned—to redness.) It thus appears that, if we are to define a similarity relation over terms in the nonobservation vocabulary, it will have to be by reference to their (direct or indirect) relations to observation terms.

But this picture might well strike one as intolerably empiricistic. It just doesn't seem to be true that the dimensions of content along which words (concepts) can be similar are reducible to the various ways in which they can be connected to observables. There is plausibly something semantically relevant that everything subsumed by the concept *UNCLE* has in common with everything subsumed by the concept *AUNT*; but a couple of hundred years of unsuccessful empiricism suggest that what they have in common is not expressible by reference to the observable properties of aunts and uncles. Similarly,

mutatis mutandis, for the similarity between the thing subsumed by the concept ICE and the thing subsumed by the concept STEAM; or between the thing subsumed by the concept THE PRESIDENT OF THE US and the thing subsumed by the concept CLEOPATRA. And so on, endlessly.

Churchland's state space story is best understood in this context. At least since *Scientific Realism and the Plasticity of Mind* (1979), he has been attracted to network semantics and inclined to think that a good semantics must make similarity, rather than identity of content, its basic theoretical notion. But he is also suspicious of the sort of empiricism which reduces all semantically relevant relations eventually to relations to observation vocabulary. On the one hand:

The semantic identity of a term derives from its specific place in the embedding network of the semantically important sentences of the language as a whole. Accordingly, if we wish to speak of sameness of meaning across languages, then we must learn to speak of terms occupying analogous places in the relevantly similar networks provided by the respective sets of semantically important sentences of the two languages at issue. (1979: 61)

However, as Churchland goes on to say a few pages later:

the aims of translation should include no fundamental interest whatever in preserving observability . . . languages, and the networks of beliefs that they embody, have an identity that transcends and can remain constant over variations in the particular sensory conduits to which they happen to be tied, and in the particular locations within the language where the sensory connections happen to be made. Accordingly, any conception of translation that ties its adequacy to the preservation of 'net empirical content' as conceived by Quine will lead to nothing but confusion. (1979: 65–6)

For Churchland, the question is thus how to free the network picture of semantics from its empiricist assumptions, and somehow to generate a robust notion of content similarity in the course of doing so. We read his paper 'Some Reductive Strategies' (1989b) as a failed attempt at this, and we'll argue that a recidivist empiricism is in fact its bottom line.

Churchland's new proposal may now be summarized: A 'Quinean' network semantics of the sort we have been discussing can be thought

of as describing a space whose dimensions correspond to observable properties, and where each expression of the object language is assigned a position in the space. To say that the concept DOG is semantically connected to the properties of barking and tail wagging is thus equivalent to saying that it occupies a position in semantic space that is (partially) identified by its value along the barkingness and tailwaggingness dimensions. Since the empiricism of the standard network proposal resides in the requirement that the dimensions of the semantic space in which the concepts are located must correspond to observable properties, all you have to do to get rid of the empiricism is abolish this requirement. What's left are semantic state spaces of arbitrary dimensions, each dimension corresponding to a parameter in terms of which the semantic theory taxonomizes object language expressions, and where similarity of content among the object language expressions is represented by adjacency relations among regions of the space.

Let's now see how Churchland proposes to develop a theory of the semantics of mental representation that accords with this conception. Rather surprisingly, Churchland's analysis starts, not with the paradigms of intentionality (propositional attitudes and concepts), but with sensations. The more or less explicit suggestion is that if we had a treatment that provided an illuminating semantical account for them, it might generalize to mental representation at large. Let us, therefore, consider how the state space story is supposed to apply to sensations, bearing in mind that it is Churchland's account of mental representation, rather than his theory of qualitative content, that is our primary concern.

The qualitative character of our sensations is commonly held to pose an especially intractable problem for any neurobiological reduction of mental states . . . and it is indeed hard to see much room for reductive purchase in the subjectively discriminable but 'objectively uncharacterizable' qualia present to consciousness. Even so, a determined attempt to find order rather than mystery in this area uncovers a significant amount of expressible information . . . Consider . . . the abstract three-dimensional 'color cube' proposed by Edwin Land . . . within which every one of the many hundreds of humanly discriminable colors occupies a unique position or small volume . . . Each axis represents the eye/brain's reconstruction of the objective

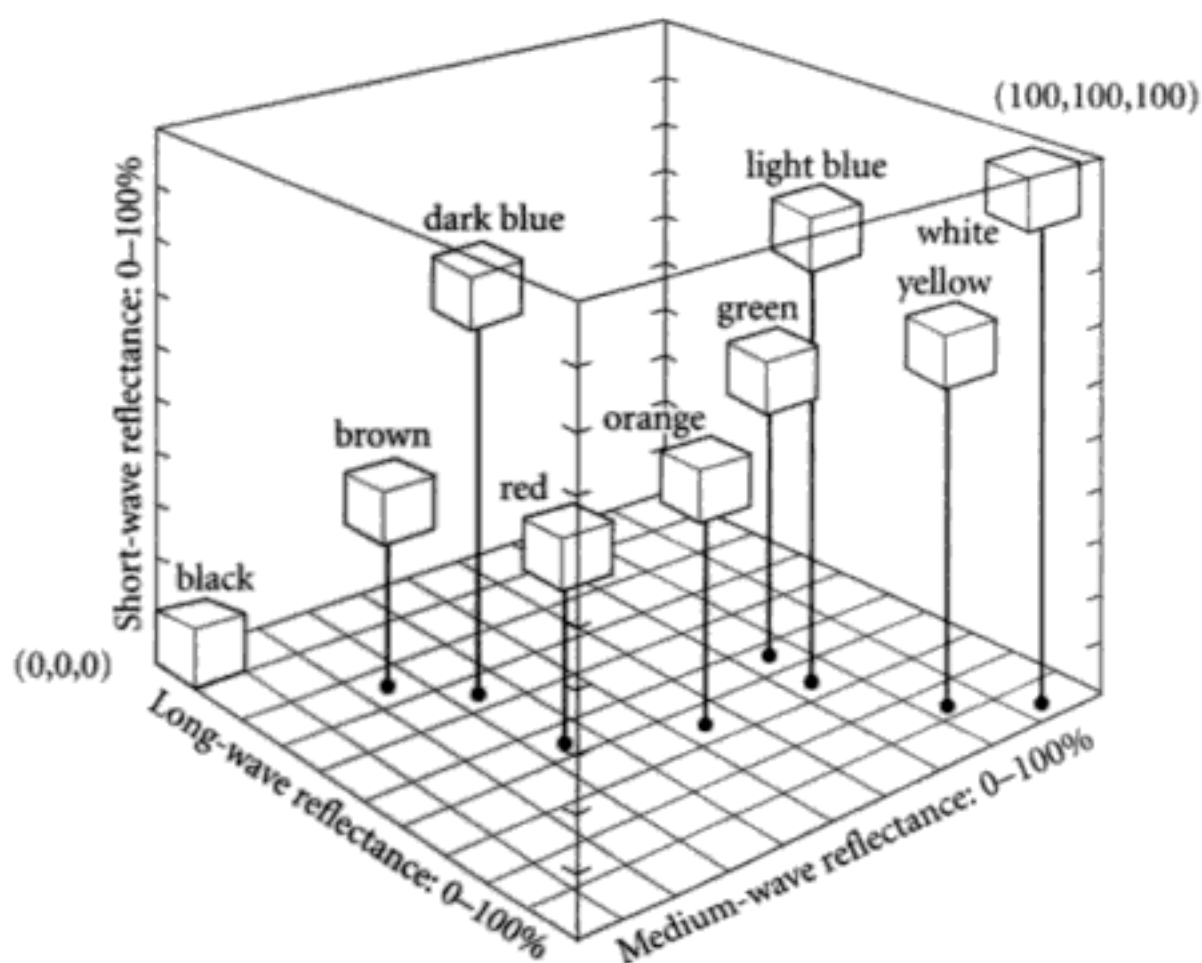


FIG. 8.1. Color state space

reflectance of the seen object at one of the three wavelengths to which our cones are selectively responsive. Two colors are closely similar just in case their state-space positions within this cube are close to one another. And two colors are dissimilar just in case their state-space positions are distant. [See Fig. 8.1.] We can even speak of the degree of the similarity, and of the dimensions along which it is reckoned. (1989*b*: 102–3)

We emphasize that Churchland views this as an account of the qualitative content of color sensations, not just of the nervous system's capacity to discriminate among colors. As he goes on to say:

In particular, it suggests an effective means of expressing the inexpressible. The 'ineffable' pink of one's current visual sensation may be richly and precisely expressible as a 95 Hz/80 Hz/80 Hz 'chord' in the relevant triune cortical system. . . . This more penetrating conceptual framework might even displace the commonsense framework as the vehicle of intersubjective description and spontaneous introspection. (1989*b*: 106)

How plausible is this story about the representation of the qualitative content of sensations? And, how close does it get us to a robust notion of content similarity in general? We'll say just a word or two about the first question, saving most of our attention for the second.

The problem of qualitative content that philosophers of mind worry about is notoriously closely connected with problems about qualia inversion. For example, it seems conceptually possible that the sensation you have when you see things that are grass-colored is 'just like' the sensation that I have when I see fire engines. If this inversion is systematic, then perhaps there is nothing—in particular, there is no behavioral consequence of our capacity to respond selectively to colors—that would tell this case apart from the normal one in which grass-colored things look the same to you as they do to me. The possibility of inverted qualia thus looks to show that behaviorism is false, and an extension of the same considerations suggests that it may show that functionalism is false too.

You might suppose that a theory of the qualitative content of sensations ought to resolve this problem. After all, it's supposed to be precisely qualitative content that gets inverted in inversion examples, and precisely the notion of identity of qualitative content that the examples render equivocal. Churchland's account of qualitative content is, however, of no help at all with these issues. The reason is that if qualia inversion makes any sense at all, it seems conceptually possible that you and I should share the state space pictured in the figure, but that the labels on your cube should be inverted with respect to the labels on mine. Notice that the reason this seems to be conceptually possible is that *the dimensions of this state space specify physical properties of visual stimuli rather than parameters of qualitative content per se*. Since the relation between the property of being a 95 Hz/80 Hz/80 Hz chord and being a sensation of ineffable pink would appear to be thoroughly contingent (or, in any event, thoroughly nonsemantic), it would seem to be conceptually possible that something should have the first property but not the second. This just is the qualia inversion problem; there appears to be no property of a sensation except its qualitative content, upon which its qualitative content is guaranteed to supervene. (In particular, there appears to be no behavioral, or functional, or neurological property upon which it is guaranteed to supervene.) So if you were worried

about the qualia problem before you read Churchland, what you should do is keep worrying.

To put this same point in an old-fashioned way, the dimensions of Churchland's state space appear to specify qualia by reference to properties they have nonessentially, and any such specification begs the inversion problem. (Or, if you think that it is 'metaphysically necessary' that color sensations have the psychophysical properties that they do, then our point is that this necessity is not engendered by any semantical connection between sensation concepts and psychophysical concepts.) You could, in consequence, know perfectly well that a certain sensation corresponds to a certain 'chord in the relevant triune cortical system' and have no idea at all of 'what it's like' to have a sensation of that kind, or, indeed, that there is anything that it is like.

The problem so far is that the dimensions in terms of which Churchland proposes to taxonomize qualia don't specify their content. Rather, they appear to taxonomize qualia by psychophysically sufficient conditions for having them. But it might be thought that this is a defect of the example, not a defect of state space semantics as such. Why not stick to the state space notion of mental representation, but add the proviso that the dimensions of the semantic space must really be semantic, not merely psychophysical; they have to taxonomize content-bearing states by their contents. Perhaps concepts (like AUNT, UNCLE, STEAM, ICE, THE PRESIDENT OF THE US, and CLEOPATRA) can be identified with positions in a state space of semantically relevant dimensions, so that similarities among these concepts could be identified with adjacencies in the state space. (The President of the US is close to Cleopatra on the politician dimension, but maybe less close on the nubile dimension.) This, as opposed to the vicissitudes of Churchland's treatment of qualia, is the issue we are really interested in.

In fact, however, we now propose to argue that this suggestion is without substance. The same problems that traditionally arose for theories of content identity also arise for this theory of content similarity, so the appearance of progress is simply an illusion. To begin with the crucial point: the state space story about content similarity actually presupposes (and therefore begs) a solution to the question of content identity.

Problems of State Space Semantics

The Individuation of Dimensions

What Churchland has on offer is the idea that two concepts are similar in content insofar as they occupy relatively similar positions in the same semantic state space. The question thus presents itself: when are S1 and S2 the same semantic state space? When, for example, is your semantic space a token of the same semantic state space type as mine?

Well, clearly a necessary condition for the identity of state spaces is the identity of their dimensions; specifically, identity of their semantic dimensions, since the current proposal is that concepts be located by reference to a space of semantically relevant properties. We are thus faced with the question when x and y are the same semantic dimensions (for example, when positions along x and y both express degrees of being a politician, or of nubility). But this is surely just the old semantic identity problem back again. If we don't know what it is for two words both to mean nubile, then we also don't know—and for the same reasons—what it is for two spaces both to have a nubility dimension. Perhaps it will be replied that semantic similarity doesn't, after all, require concepts to be adjacent in the *very same* state space; perhaps occupying corresponding positions in *similar* state spaces will do. That a regress has now appeared is, we trust, entirely obvious.

It's worth getting clear on what has gone wrong. The old (empiricist) version of network semantics had a story about the identification of the dimensions by reference to which it did its taxonomizing; they were to express observable properties, and an externalist (for example, a causal) theory of some kind was to explicate the relation between observable properties and terms in the observation vocabulary. In particular, that relation was assumed to be specifiable independent of the interpretation of the rest of the vocabulary. However, as we've seen, Churchland's proposal comes down to the idea that the dimensions of semantic state space *don't* generally correspond to observable properties after all; they can correspond to whatever properties the brain may represent. This avoids empiricism, all right, but it begs the question how identity of state spaces is itself to be

determined. On the one hand, we are assuming that dimensions of semantic state spaces can express whatever properties the brain can express. And, on the other hand, we don't have and can't assume any identity criterion for dimensions that express other than observable properties. And, on the last hand, to take such a criterion for granted would just be to beg the semantic identity problem.

To repeat: We have a robust notion of semantic similarity only if we have a criterion for the identity of state spaces. We have a criterion for the identity of state spaces only if we have a criterion for identity of dimensions of state spaces. And we have a (nonempiricist) criterion for the identity of dimensions of state spaces only if we have a criterion of 'property expressed by a dimension of a state space' that works for arbitrary properties, not just for observable properties. But a criterion for 'property expressed' that works for arbitrary properties is just a criterion for identity of meaning. So Churchland's proposal for a robust theory of content *similarity* fails to avoid the problem of robust content *identity*. (And, of course, fails to solve it.)

In our book *Holism, a Shopper's Guide* (1992), we offer it as a plausible methodological principle that you can't have a robust notion of content similarity (one that applies across languages, across minds, or across theories) unless you have a correspondingly robust notion of content identity. Churchland's state space semantics provides a graphic illustration of how this principle applies. His explication of an interpersonal notion of content similarity as proximity in semantic state space presupposes an interpersonal notion of identity for the semantic spaces themselves, a notion that Churchland leaves entirely without explication. In consequence, if you're worried about how concepts can be robust, Churchland's state space semantics provides no illumination at all.

We're claiming, in effect, that Churchland has confused himself by taking the labels on the semantic dimensions for granted. The label on a dimension says how positions along the dimension are to be interpreted; for example, it says that they're to be interpreted as expressing degrees of F-ness. To label a dimension as the F-ness dimension is thus to invite the question 'In virtue of what do the values of this dimension express degrees of F-ness rather than, say, degrees of G-ness?' (Equivalently, for these purposes, 'What makes it the case that a dimension in your state space expresses the same prop-

erty F that some dimension in mine does?') Patently, a semantic theory mustn't beg this sort of question on pain of assuming the very concepts it is supposed to explicate.

Cognitive scientists are forever getting themselves into trouble in this way; it's a fallacy that is particularly endemic among connectionists. Connectionists draw diagrams in which the label on a node tells you what the intentional interpretation of the excitation of the node is supposed to be. But no theory is offered to explain why a node gets the label that it does; it's just semantics by stipulation. Churchland makes exactly this mistake, only it's the dimension labels rather than the node labels that he stipulates. This, however, is actually a distinction without a difference since stipulating a semantic interpretation for the dimensions *just is* stipulating semantic interpretations for points (regions, and the like) in the space they define. This fact is obscured in Figure 8.1 because the labels that are provided for points in that semantic space (namely, the color labels 'brown', 'dark blue', and so on) are not actually the ones implied by the labels on the axes. The labeling of the axes implies that points in the space are ordered triples corresponding to values of short-, medium-, and long-wave reflectances. By contrast, the node labels represent not the semantics of the state space but Land's empirical proposal about how the qualitative character of sensations varies as a function of psychophysical properties of light.

This slight tension in Churchland's notation turns out, upon reflection, to be the tip of a substantial iceberg. Consider the question: How does Churchland decide what gets represented by *dimensions* of state spaces and what gets represented by *regions* in the state spaces that the dimensions define? What decides, for example, that 'brown' and 'dark blue' correspond to regions rather than dimensions? There is, after all, nothing obviously wrong with taxonomizing bananas according to their degree of brownness or oceans according to their degree of dark blueness, and there are presumably lots of cases where one's concept of an X includes information about the color of Xs—blondes for example.

As we've seen, one principled way of making this decision might be to stipulate that the dimensions of state spaces express, always and only, psychophysical parameters; and, this may well be what Churchland has in mind. If so, however, what he is offering isn't a

semantics at all. For, a semantics taxonomizes mental states by their contents, not by their causes. Churchland may be assuming that if a stimulus has such and such psychophysical properties, then if the brain responds to the stimulus, the brain thereby represents the stimulus as having those psychophysical properties. But the brain represents red things as red, not as reflecting light of such and such wavelength; and it represents aunts as aunts, not as possessing whatever psychophysical properties we employ for purposes of aunt-detection (assuming, indeed, that there are such properties). Psychophysics would be a lot easier were this not so; we could do it by introspection.

Churchland's state spaces thus vacillate between being psychophysical spaces and being semantic spaces. Correspondingly—and this point really is essential—proximity in state space means quite different things in the two interpretations of Churchland's theory. According to the semantic reading, proximity expresses the similarity of the content of mental states; according to the psychophysical reading, it expresses the similarity, under physical description, of the (proximal) stimuli that elicit the sensations. Apparently Churchland hasn't decided which sort of similarity he has in mind, so that there's a crucial respect in which he hasn't decided what his space state theory is a theory of.

That Churchland really has confused the semantical with the psychophysical enterprise is suggested by revealing hesitations like his suggestion that state spaces offer 'an *alternative to, or potential reduction of*, the familiar Chomskian picture' (1989a: 109; our emphasis). One or the other, perhaps, but surely not both. An *alternative* to the Chomskian picture would be a new story about what a native speaker knows about his language; that is, a new story about the intentional content of the native speaker's knowledge of his language. *Qua* alternative it would, by definition, be in competition with Chomsky's picture. A *reduction*, by contrast, would presumably be a story about the neural format in which what the native speaker knows about his language is coded in his brain. By definition, a reduction is not in competition with the theory that it reduces. The semantic theory and the reductive theory are in quite different lines of work, and nothing can do both jobs.

The Analytic/Synthetic Distinction

As you might expect, all the other standard worries about content identity now come trooping back. Consider the a/s question itself. It's analytic (let's say) that dogs are animals; it's not analytic (let's say) that they typically have wet noses. Consequently, to change to Churchland's notation, the space in which the concept DOG is resident must have a dimension corresponding to the property of being an animal, but it needn't have a dimension corresponding to the property of typically having a wet nose. (In fact, it had better not have such a dimension if the location of an item in the space is to predict its behavior in modal inferences.) This is to say that it is constitutive of the concept DOG that it subsumes only animals, but not constitutive of the concept DOG that it subsumes only things whose noses are typically wet. Problem: If you are convinced that there is no principled way of drawing an a/s distinction, to what principle will you appeal to distinguish the dimensions that are relevant to defining semantic spaces from the dimensions that aren't? You can have it as the analytic/synthetic problem as the problem of saying which inferences are constitutive of a concept, or you can have it as the problem of saying what makes something a bona fide dimension of semantic state space (as *per* Churchland). The point is that it's the same problem whichever way you choose to have it; in particular, it doesn't go away when you start to think of semantic relations among concepts as 'metric'.

Collateral Information

If it is to turn out that your concept DOG occupies much the same region of your semantic space that my concept DOG occupies in mine, then it had better be that your semantic space has whatever dimensions mine has; as we've seen, this raises the question of how the dimensions of semantic state space are to be identified/individuated, which would seem to be the content identity problem all over again. Well, suppose the problem of individuating dimensions is somehow solved; it's still not guaranteed that your DOG concept and my DOG concept will turn out similar. The problem is that I know a lot of

things about dogs that you don't. (And, of course, vice versa.) For example, I know that I once had a dog named Spot, and that my grandmother is allergic to Dalmatians, and that there aren't any dogs in the room in which I'm writing this. But you wouldn't have known any of those things about dogs if I hadn't just told them to you. (Nor is there any particular reason why the various beliefs about dogs that we fail to share must in general be token-reflexive. If you're a dog buff, and I'm not, you have very many standing beliefs about dogs that I don't have and that I don't have any desire to acquire.)

The point is that if a semantics recognizes dimensions of state space corresponding to all the properties of dogs about which our beliefs differ, then even assuming that your state space has exactly the same dimensions as mine, the location of the dog concepts in our respective spaces is likely to turn out to be quite significantly different. This should all be sounding like old news; it's just the worry, familiar from attempts to construct a notion of content identity, that a lot of what anybody knows about dogs counts as idiosyncratic; it's 'collateral information', the sort of thing that Frege says belongs to psychology rather than semantics. If we are to have a notion of meanings as shared and public property, a robust notion of meaning, we must somehow abstract from this idiosyncratic variation. This holds just as much for similarity of meaning as it does for identity of meaning. Or, if it doesn't, an argument is needed to show why it doesn't. We're not aware that there's any such argument around.

If, in short, all dimensions count in determining positions in state space, then even if everybody's state space has exactly the same dimensions as everybody else's, the collateral information problem makes it quite unlikely that anybody's concepts will turn out similar to anyone else's. One way out would be to assume that the dimensions of state space are weighted; that agreement along some dimensions counts more for conceptual similarity than agreement along other dimensions does. In traditional theories of content identity, this weighting of dimensions was accomplished precisely by appealing to the a/s distinction: it matters that your concepts have the same values as mine on dimensions that correspond to properties that figure in the analysis of the concepts; with respect to the other dimensions, the concepts are allowed to vary idiosyncratically. Once again, nothing changes when the topic changes to content similarity. Our concepts

are similar if they have similar locations along certain dimensions, whether or not they have similar locations along other dimensions. The collateral information problem is to find a principled way of deciding which dimensions count a lot, which ones count a little, and which ones don't count at all. And nobody has the slightest idea of how to do this without invoking an a/s distinction.

We hope the moral will now be clear. What Churchland's got is a dilemma: it may be that he doesn't require that the dimensions of his state space correspond to properties of the contents of the mental states (objects, events) that they taxonomize. In that case, he isn't doing semantics at all. He's doing, as it might be, psychophysics, and we have no quarrel with anything he says except that he is not really entitled to describe what he has on offer as a theory about how 'the brain *represents* various aspects of reality by a position in a suitable state space' (our emphasis). If, on the other hand, Churchland is taking the talk about neural representation seriously, his move to state spaces leaves all the old problems about content identity still to be solved. We think that probably, given these options, Churchland would prefer the first. He is, as we remarked above, very much inclined to be an eliminativist about intentional/semantic properties when metaphysical push comes to shove. An eliminativist doesn't, however, need a notion of semantic similarity any more than he needs a notion of semantic identity. An eliminativist doesn't want to reconstruct semantical discourse; he wants to change the topic.

But couldn't Churchland somehow contrive to have it both ways; couldn't he somehow work it out that state spaces do semantics and psychophysics at the same time? Sure he could. What he has to do is just assume the 'empiricist principle' that all our concepts are functions (possibly Boolean functions, but more likely statistical functions) of our psychophysical concepts. As far as we can tell, this is what connectionists, Churchland included, actually do assume (though, understandably, they aren't eager to put it that way). Here, for example, is Churchland doing his level best to make a virtue of this necessity. Anyone who wants to understand which epistemological assumptions underlie the connectionist version of state space semantics should pay especially close attention to the passage that starts 'it may be that . . .'

the activation-vector spaces that a matured brain has generated, and the prototypes they embody, can encompass far more than the simple sensory types such as phonemes, colors, smells, tastes, faces, and so forth. Given high-dimensional spaces, which the brain has in abundance, those spaces and the prototypes they embody can encompass categories of great complexity, generality, and abstraction . . . such as harmonic oscillator, projectile, traveling wave, Samba, twelve-bar blues, democratic election, six-course dinner [etc.] . . . It may be that the input dimensions that feed into such abstract spaces will themselves often [*sic*] have to be the expression of some earlier level of processing, but that is no problem. The networks under discussion are hierarchically arranged to do precisely this as a matter of course. In principle, then, it is no harder for such a system to represent types of processes, procedures, and techniques than to represent the 'simple' sensory qualities. From the point of view of the brain, these are just more high-dimensional vectors. (1989a: 191)

So, apparently, the picture is that neural representations form a hierarchy, with sensory representations at the bottom and the rest somehow abstracted from them. Take-home exercise: What is there in this understanding of the relation between abstract and sensory concepts that Hume would not have endorsed?

The reason that assuming the empiricist principle is helpful in this context is that it provides a robust identity criterion for semantic spaces: If all concepts are ultimately sensory concepts, then every concept is a location in a semantic space whose dimensions express sensory properties. So two organisms share a semantic space if they have their sensory transducers in common. The reason why assuming the empiricist principle is nevertheless to be avoided, however, is that it isn't true. Most concepts are not Boolean or statistical functions of psychophysical or sensory concepts, and most concept learning is not either Boolean or statistical inference. *A fortiori*, one's concepts and theories are not to be identified with points in a space whose dimensions express sensory or psychophysical properties or constructs thereof. You might have thought this was the one thing philosophers would have learned from 200 years or so of epistemological bad weather. But, no: the smoke clears and the landscape is revealed as having hardly changed at all.

That Churchland's account of content similarity is question-begging doesn't, of course, constitute a proof that a robust notion of

content similarity can't compensate holistic semantics for its lack of a robust notion of content identity. Somebody might come up with a robust notion of content similarity that's not question-begging by early tomorrow afternoon (though one would, perhaps, be well advised not to hold one's breath). But what's not a proof may nevertheless serve as an object lesson. The current situation is that nobody has any idea what content similarity is, just as nobody has any idea what content identity is—and, as we remarked at the outset, for much the same reasons.

Appendix 8.1 Reply to Churchland's (1993) Reply

Stage One Churchland—the Churchland whose papers we discussed in *Holism* (1992)—identified concepts with locations in a state space whose dimensions correspond to sensory parameters. Nothing distinguishes this from traditional, naïve empiricism except the terminology: where naïve empiricists speak of *sensory* properties of objects (that is, the properties of objects to which sense organs are responsive), a Stage One Churchland connectionist speaks of *transducible* properties of objects (that is, properties to which transducers are responsive).

There is a notorious (Berkeleyan) objection to naïve empiricism, and, unsurprisingly, it holds against Stage One Churchland too: Most concepts aren't constructions out of sensory concepts. Justice and happiness, for example, can't be imaged, because, to put it crudely, what the just share *qua* just and the happy share *qua* happy can't be specified by reference to their sensory (= transducer-detectable) properties. You can image a triangle, of course, but you can't have an image of triangularity as such; what triangles have in common *qua* triangles also isn't something that can be expressed in sensory terms. So far, there is nothing in Stage One Churchland except what Locke proposed and Berkeley refuted. Or so we claim.¹

¹ Churchland, however, thinks the analogy between naïve empiricism and Stage One is a bad one. Whereas naïve empiricists thought of complex concepts as *constructed* out of sensory primitives, connectionists think of complex concepts as 'filtered' out of the background of correlated properties. This is, however, a distinction without a difference. Suppose *glub* is the concept of something round and red. Then you can think of learning the concept as constructing *glub* from instances of the concepts of redness and roundness; or you can think of learning *glub* as filtering red-roundness out of a background of irrelevant concomitant stimulation. Suppose you sort a pile of rocks into the red ones and the brown ones. You can think of the process

There is, to be sure (and as we remarked in *Holism*), one difference between the sort of concept learning that Churchland has in mind and the sort of concept learning that naïve empiricists contemplated: Churchland is not just a meaning holist, but a connectionist as well. These days, as he says, a connectionist is the natural thing for a meaning holist to be. And, whereas naïve empiricists supposed concepts to be something like *Boolean* functions of sensory primitives, connectionists think of concepts as something like *statistical* functions of sensory primitives. Churchland finds this ability of neural nets to detect statistical relations 'marvellous', 'almost magic'. But in fact, as far as anybody knows, what networks do (analogically) is much the same sort of thing that familiar statistical feature extraction procedures do (digitally). Networks are about as magical as factor analysis, and a lot slower.

As we pointed out in *Holism*, Stage One Churchland has a dilemma. If the dimensions of semantic state space are stipulated to be sensory, the state space theory of concepts reduces to naïve empiricism and is thus born pre-refuted. The alternative is to liberalize the allowable dimensions. The concept *justice* can't be a location in a space of *sensory* dimensions; but, trivially, it could be a location in a space one of whose dimensions is *justice*. Or, to say the same thing without the intimidating talk of spaces and vectors: you can't define 'justice' in terms of 'red', 'triangular', and such; but, of course, you can define it in terms of 'justice'. The options Churchland has on offer are therefore to do without *justice* or to resign yourself to empiricism. It was, of course, precisely this line of thought that led Hume, in some of his moods, to deny that we have concepts like justice, triangle, cause, etc.

Stage Two Churchland—the Churchland of the present paper—thinks he has a way out: let the dimensions of semantic space continue to be sensory but what makes a neural representation a concept of kittens isn't its sensory constitution but its 'downstream' role in certain inferences. The 'prototype vector has rich causal and computational sequelae of its own . . . because of its own transformational significance . . . downstream' (1989c: 276). This is Hume's solution to Berkeley's problem about abstract ideas: The concept of a triangle can't be an image, but it might be an image that is *appropriately employed in reasoning about triangles*. Having the abstract concept is having an image of some triangle or other, together with rules that tell you what to ignore when you use the image to think about triangles. Hume's empiricism thus distinguishes the concept of a triangle from the concept of a prototypical triangle, and identifies the former with the latter together with an inferential regime.

as selecting the red ones, or you can think of the process as filtering out the brown ones. To prefer one way of speaking over another is to prefer your cup half-empty to your cup half-full.

But this leaves wide open Quine's question how the inferential regimes are themselves to be individuated. Stage Two Churchland offers *no* proposal for answering this question. Churchland says nothing about how Paul and Patricia, whose kitten prototypes are associated with *different* inferential regimes, could nevertheless both have the concept KITTEN. We thought that he thought that state space semantics was going to answer this sort of question by providing a viable notion of content similarity, thereby relieving holism of its chief embarrassment. Apparently, however, this project has now been quietly dropped.

All at Sea in Semantic Space: Churchland on Meaning Similarity

IT is old news that the identity conditions for contents, meanings, and the like are problematic, perhaps because there cannot be identity conditions for meanings unless there is an analytic/synthetic distinction, and there isn't. We aren't proposing to consider whether the notion of content identity can be made metaphysically respectable. But we want to discuss a proposal about how to save semantics if it does turn out that 'same meaning' can't be reconstructed; it is one that is currently popular in both the philosophical and the cognitive science literature.¹

The idea is that the purposes semantic theories are supposed to serve do not actually require principles of content *individuation*; all they require is a notion of content *similarity*. On such a view, similarity of meaning is primitive relative to sameness of meaning, and the latter might be of interest only as a limiting case of the former; perhaps, indeed, a limiting case that is never achieved in practice. Meaning similarity is real and robust, there is a lot of it around, and it is metaphysically prior to synonymy. So the story goes.

We have two points to make here. First, even if a notion of content similarity were on offer, it would arguably be of no use for doing the kinds of things that a theory of meaning ought to do. Quite generally, the kinds of explanations that semantic theories are supposed to

¹ Texts where this sort of suggestion is endorsed are ubiquitous. For some examples: Harman (1973); Block (1986: 615–78); Sloman and Rips (1998); Stich (1981: 153–206); Jackendoff (1988: 81–97).

give would not survive substituting a similarity-based notion of content for an identity-based notion. This is unaccidental; it is the result of ways in which content similarity and content identity are intrinsically different kinds of relations. Second, we shall argue that, even if a similarity-based notion of content could in principle replace an identity-based notion *salve* the explanatory power of semantic theories, that would not be much cause for rejoicing. For, as a matter of fact, nobody has the slightest idea how to construct the required notion of content similarity; and the reasons for this are much the same, *mutatis mutandis*, as the reasons why nobody has the slightest idea how to construct a perspicuous notion of content identity. Until recently, we would have thought this claim was untendentious. But Paul M. Churchland, in a number of articles over the last decade, has proposed what he says is a schema for such a construction (1989b: 77–110; 1993: 667–72; 1996: 278–83; 1998: 5–32; see also Fodor and Lepore 1992, ch. 6; 1993d: 679–83). That he is quite wrong to say this has been the burden of several papers of ours, to which Churchland has now likewise replied. Our view is that his reply is an egregious *ignoratio elenchi*. The second part of the present discussion is our attempt to make clear why it is.

Why Content Similarity is Probably of No Use

A question that isn't asked nearly as often as it should be is this: If you had a tenable notion of similarity of meaning, what exactly would you do with it? We suspect that friends of similarity-based content more or less take for granted that, given such a notion, they could just plug it in wherever semantic theories have hitherto availed themselves of synonymy, content identity, intentional equivalence, and the like. It is (we continue to suspect) because the friends of meaning similarity assume this that they rarely bother to show how a semantics built around content similarity could do anything like the explanatory work that semantic theories are supposed to do.² In fact,

² Arguing for this assumption is a burden which semantic *eliminativists*, unlike the friends of content similarity, are of course not required to discharge. There are people around, including Churchland himself in some of his moods, who apparently can't decide whether their metaphysical bottom line is realism about content similarity

a plausible *prima facie* case can be made against this assumption; so, anyhow, we are about to argue.

We can think of four main purposes for which one might want a theory of meaning and for which such theories have historically been employed. We don't deny that there may be other uses to which they could be put; but we do think it's plausible that, if a theory of meaning failed all of the following, it would not be worth having.

Satisfaction conditions. A semantic theory for a language ought to assign satisfaction conditions to its semantically evaluable expressions; likewise, *mutatis mutandis*, for a theory of the intentional contents of semantically evaluable mental states.³ Moreover, it should turn out that sentences that we are all quite certain are true *are* true under the preferred assignments of satisfaction conditions. For example, *ceteris paribus* there would be something badly wrong with a semantic theory that assigned truth conditions to 'Sparrows are birds', 'Water is wet', 'George Washington was president', and so on, according to which any of them expresses a falsehood.

Compositionality. A semantic theory should make clear how the satisfaction conditions for complex linguistic expressions are determined by the satisfaction conditions of their grammatical parts, thereby explaining why natural languages are typically productive and systematic. Likewise, a semantics for propositional attitudes should make clear how the satisfaction conditions for beliefs, desires, and the like are determined by those of their constituent concepts, thereby explaining why minds are typically productive and systematic.

Translation. Pretheoretic intuition has it that meaning is what good translations preserve. A semantic theory should provide a notion of meaning according to which this turns out true.

Intentional explanation. A semantic theory should reconstruct a notion of content that is adequate to the purposes of intentional (for example, belief/desire) explanation.

or skepticism about meaning. Our discussion is only concerned with the former doctrine.

³ We assume that linguistic expressions and propositional attitudes are both semantically evaluable, and we are neutral as to which, if either, is ontologically prior. We will generally go back and forth between the semantics of languages and the semantics of mental states as expository convenience suggests.

We are about to argue that, quite likely, a similarity-based theory of meaning would not do any of these four kinds of things; not even if one assumes, *qua* semantic realist, that an identity-based theory of meaning would do the work if only the individuation problem for contents could be solved.

Satisfaction Conditions

Consider, to begin with, any statement (or sentence or thought) you like which you are prepared to believe is atomic, has the logical form Fa , and is true. Nothing much except ease of exposition turns on how notions like 'logical form' and 'atomic statement' are deployed; you choose. (In fact, it doesn't even matter all that much how you construe the notion 'true'; we shall return to this.) Take, for example, the sentence (S1):

(S1) Nixon is dead.

By our lights,

- (1) (S1) is certainly true.
- (2) 'Nixon is dead (at t)' is true if and only if Nixon is dead (at t).⁴
- (3) No semantic theory (for English) could be adequate unless it entails (2) and is consistent with (1).

Query: What would a content-similarity-based semantics say about the truth conditions of (S1)? Well, we don't know because, of course, there aren't really any such theories on offer. But we guess the story would go something like this. Think how an old-fashioned, identity-based semantic theory goes about entailing (2). Roughly, it assigns the individual Nixon to the singular term 'Nixon'; it assigns the property of *being dead* to the expression 'is dead', and it interprets sentences of the logical form Fa to mean that the individual designated by ' a ' has the property expressed by ' F '. Well, likewise, *mutatis mutandis*, according to the new dispensation; except that, for example, we assign to the expression 'is dead' not the property of being dead, but a *range of properties all of which are similar to being dead*,

⁴ From here on, we ignore issues about context-sensitivity; in particular, we won't worry about implicit indexicals.

and we assign to 'Nixon' a range of individuals all of whom are similar to Nixon.⁵

But, clearly, to proceed in this fashion is simply to give up the hope of assigning satisfaction conditions to formulas in a way that makes, for example, sentences like (2) come out true. Thus, we suppose that being comatose on one's deathbed is pretty similar to being dead; but it is not the case that, if Nixon is comatose on his deathbed at *t*, then Nixon is dead at *t*. If you don't believe us, ask Dr Kevorkian.

We take it to be just self-evident that properties constructed out of semantic similarity will not do what notions like 'expresses', 'denotes', 'means', and so on are supposed to do; in statements like 'Nixon' denotes Nixon, 'is dead' expresses the property of being dead, 'is dead' means *is dead*, and the like. The problem is simple enough: if two expressions denote, express, or mean the same, then they are coextensive; but if two expressions only denote, express, or mean *something similar*, they needn't be and typically aren't. And it is at the heart of the standard way of assigning satisfaction conditions that what replaces 'F' on the left-hand side of instances of the schema 'Fa is true if and only if *a* is F' must be at least coextensive with what replaces it on the right-hand side. The moral seems to be that part of the price of switching from a content-identity-based semantics to a content-similarity-based semantics is giving up the idea that theories of meaning specify the pretheoretically correct conditions of semantic evaluation of thoughts, sentences, and the like.⁶

⁵ Perhaps, however, names express individual concepts, in which case 'Nixon' is assigned some property similar to the one that 'Nixon' expresses. We are not at all clear what a similarity-based semantics should say about singular terms. Since the options for predicates seem a bit clearer, from here on we will generally run the discussion on them.

⁶ We are assuming that the project is to replace a semantics based on intentional identity with one based on intentional similarity, but we are also assuming that the usual notion of identity of *extension* survives. That is, it is not part of the story that the revisionist semantics would appeal to some notion of similarity of *extension* where traditional semantics adverts to extensional identity. (As far as we know, none of the proposals for a similarity-based reconstruction of the notion of intention on offer in the literature does call for a similarity-based reconstruction of the notion of extension: either the issue is not discussed, or the conservative notion of extension is assumed. See, e.g., Putnam (1975: 215–71).)

In fact, we can't imagine how a similarity-based notion of extension would go. The problem is that the similarity of two extensions would presumably have to be grounded not in their *degree of overlap*, but in the *similarity of the individuals* that

A last consideration under this general head: you might suppose that you could use a similarity-based semantics to assign satisfaction conditions in something like the traditional way if you are prepared to give up an absolute notion of *truth* in favor of some correspondingly approximate notion. On this view, for example, the right truth sentence for (S1) might be not (2) but (4).

- (4) 'Nixon is dead' is *similar to being true* (that is, it is roughly true) if and only if Nixon is similar to being dead.

We don't actually know of anybody into meaning similarity who clearly endorses this, nor are we sure what, exactly, such a proposal would amount to. But whatever (4) means, there are surely counter-instances to the schema *formulas of the form Fa are roughly true if and only if a has some property similar to F*. The basic problem is that 'similar to', unlike 'roughly true', is, of necessity, relativized to respects. Thus, it is true (we suppose) that dogs are similar to cats; but it is not even roughly true that dogs are cats; or that our cat is a dog . . . and so on. The equivalence also fails in the other direction. It is roughly true that you cannot be the Pope if you were born in the Bronx. But being the Pope and not being born in the Bronx are not at all the same sort of thing. Both of the present authors are instances to the contrary.

Compositionality

Compositionality is the idea that, in the typical case when syntactically complex expressions(/thoughts) are semantically evaluable, their satisfaction conditions are determined by the satisfaction conditions of their syntactic constituents. And it is generally agreed that the crucial condition that must be satisfied in order that

belong to their extensions; and who knows how this notion is to be explicated? Thus, for example, the extension of 'male college senior' is, presumably, 'more similar' to the extension of 'female college senior' than it is to the extension of 'rock' or of 'prime number'. But none of these sets overlaps at all. By contrast, the set {male truck drivers} is less similar to the set {male college students} than it is to the set {female college students}, though the first and second probably overlap and the second and third do not.

We shall, however, briefly consider a case of a similarity-based semantic theory which proposes to dispense not only with identity of meaning but also with identity of truth value. See immediately below.

compositionality should hold for a class of expressions to be compositional is that the satisfaction conditions of their constituents should be *context-independent*. So, the fact that 'brown dog', 'green dog', 'brown cat', and 'yellow cat' are all compositional in English is part and parcel of the facts that 'brown' means the same in the environment '... dog' that it does in the environment '... cat'; 'dog' means the same in the environment 'brown ...' that it does in the environment 'green ...'; and so on. Likewise, *mutatis mutandis*, for 'brown poodle dog' and the like. That compositionality requires context-independence is, to repeat, the consensus view, and we simply take it for granted in what follows.

Now, the claim that the meaning of a constituent is context-independent is essentially the claim that schema (S2) is valid:

- (S2) If *m* is part⁷ of the meaning of '*a*' and '*a*' is a constituent of '*b*', then *m* is part of the meaning of '*b*'.

The point of present interest, however, is that (S3)—which is what you get if you substitute 'is similar to part of the meaning of' for 'is part of the meaning of' throughout (S2)—is surely false:

- (S3) If *m* is similar to part of the meaning of '*a*' and '*a*' is a constituent of '*b*', then *m* is similar to part of the meaning of '*b*'.

So, presumably, the meaning of 'cat' is similar to part of the meaning of 'The leopard is on the mat' (cats and leopards are both felines) and part of the meaning of 'The leopard is on the mat' is similar to part of the meaning of 'The explosion caused a lot of damage' (leopards and explosions are both dangerous). But the meaning of 'cat' is not similar to any part of the meaning of 'The explosion caused a lot of damage'. (However, see the appendix to this chapter.)

The main argument is, we think, simple and obvious: the standard explication of the compositionality of content depends crucially on assuming that the semantic property of a constituent which is preserved when it is embedded in its host is content identity, not just content similarity.

⁷ Reading 'part' as 'improper part'. So, in particular, the whole meaning of a syntactically primitive expression counts as 'part of its meaning and as part of the meaning of any complex expression of which it is a constituent'.

Translation

The relation *translates*, like the relation 'is similar to the meaning of', is plausibly intransitive. For this reason, many philosophers think that, whatever the flaws of similarity-based theories of content may be, they are at least likely to be useful in accounts of translation. And, indeed, there are philosophers who think this is the *only* work a semantic theory can legitimately be asked to do because facts about translation are the only bona fide semantic facts there are.

We think, however, that this line of thought is extremely dubious, at least for the case of infinite languages (namely, languages of infinite expressive power), which we suppose to be the only case worth caring about. Suppose that a translation theory is a finitely specifiable, computable function from the sentences of L_1 to the sentences of L_2 , under which similarity of meaning is preserved. Then presumably there cannot be such a theory unless L_1 and L_2 are compositional. The basic idea is pretty obvious. Suppose the languages are infinite and the translation theory is finite. Then presumably the translation theory works by first listing the translation relations among the primitive bases of the two languages (it says that 'chat' means the same as 'cat' and that 'chien' means the same as 'dog', and so forth) and then providing recursive procedures for constructing the translation relations among the infinitely many complex expressions. But, we suppose, this will work only where, in each of the two languages, the semantics of the complex expressions is determined in a regular way by the semantics of the primitive basis; that is, only if the languages are semantically compositional.⁸ If this line of thought is right,

⁸ But what if translation turns out to be a *syntactic* relation? That is, what if the translation of L_1 expressions into L_2 expressions is fully determined given their respective syntactic structural descriptions (including, of course, their lexical inventories; see, e.g., Schiffer (1987); Fodor (1990, ch. 7); Lepore (1996: 41–60)). This is perfectly all right with us, but notice that it is no help to somebody who says that translations *ipso facto* preserve similarity of meaning. The claim that syntax *determines* translation is not at all the same as claiming that syntax is what (good) translation *preserves*; what translation preserves, according to similarity-based semantics, is similarity of content. Well, if it is to do so in the case of translation between productive languages, infinitely many meaning-similarity relations among complex expressions must be projected from finitely many meaning-similarity relations among primitive expressions. But, as we remarked in the text, it looks like meaning similarity is generally *not* preserved under the recursive processes that construct complex expressions from simpler ones; in general, the semantic property such processes can preserve is content

then the arguments that hold against similarity-based compositional theories (see above) are also arguments against similarity-based translation theories.

Intentional Explanations

Paradigm intentional explanations appeal to principles that control the relations (rational, or causal, or both) between propositional attitudes with identical or overlapping contents. Thus, the 'practical syllogism' is supposed to be captured by some such schema as (S4):

- (S4) If *a*(gent) desires that *P* and believes that *not-P unless Q* (where *a* believes that *Q* is contingent on an action that he believes himself able to perform), then, *ceteris paribus*, *a* tries to bring it about that *Q*.

Prima facie, (S4) is sound only if the formulas that substitute for *P* are identical or synonymous throughout. (It is plausible that weaker relations like, for example, logical equivalence are insufficient even when they preserve the extensions or the relata.) Our point is that (S4) is patently unsound if substitutions for the schematic variables preserve only similarity of meaning.

Being a zircon is similar to being a diamond; zircons are manufactured to ensure that this is so. But Marilyn Monroe did not think that zircons are a girl's best friend, and she was right not to think so. Likewise, Château Rothschild is not *all that* different from Château Plonk; but there are many who are prepared to pay for the one what they would not pay for the other. Likewise, you may have a strong preference as to which twin you marry, even if the similarity between the twins is notable. And so on. None of these asymmetries of preference need indicate a palate that is excessively refined. To the contrary, the mediation of responses that discriminate sharply between similar situations is a fair part of what cognition is for. A landscape

identity. If this is right, a *syntactic* translation procedure could *not* be expected to preserve the *semantic* properties of expressions in the languages between which it operates if both languages are productive and the only semantic property defined for their expressions is meaning similarity.

without a ravening tiger may be quite similar to a landscape with one. Tigers make a living out of that; it is part of the story about why they have stripes. It may thus matter a lot which of two quite similar landscapes you are in. Your cognitive mechanisms are there in large part to help you to find out.

The underlying principle in all this is entirely obvious:

If x and y are identical, then they are identical *tout court*. But if x and y are (merely) similar, then there is always a *respect* in which x and y are similar and lots of respects in which they aren't; and it can matter to their rational or causal relations which respects are which.

Here is another way to make much the same point. Presumably, $F = G \rightarrow \Box (F = G)$. But, of course, the corresponding conditional for similarity is invalid. The result is that the counterfactuals that identity statements guarantee are not, in general, consequences of the corresponding similarity statements. We assume that 'P explains Q' is true only where 'if P, then Q' does support appropriate counterfactuals. All this being so, it is hardly surprising that, in general, you cannot replace content identity with content similarity in an intentional theory *salve* explanatory power. There may be some way for a similarity-based semantics to save the kind of intentional explanations that the practical syllogism provides; but we do not ourselves know how; and we have seen no suggestions in the literature.

We end this section by remarking that it's our guess that nothing we have said so far would move Churchland in the least. What he would reply, we continue to guess, is that we have failed to realize how very radical his revisionist proposals actually are: they are intended to abstract not only from the traditional idea that the central semantic construct is content identity, but also from the traditional inventory of problems that semantic theories are supposed to solve. Well, so be it. But we, in turn, are not moved to exchange the old way of doing semantics for a new way on the grounds that the latter, though it will not solve what used to be thought the main problems about meaning, does promise to solve some new problems that it is unfortunately not now possible to enumerate. A pig in a poke is what that sounds like to us.

Why Churchland has no Notion of Content Similarity

Thus far our topic has been what role a similarity-based notion of meaning might (or might not) be able to play in a semantic theory. Suppose, for the sake of the argument, that our doubts about this are misplaced and that an account of meaning similarity would be a very good thing to have. We turn to our second topic, which is why we do not think Churchland has got one.

This requires a bit of background. Let's start with the notion of a semantic vector space since it plays a central role in Churchland's proposal. The basic idea is illustrated by Figure 9.1. This figure is supposed to provide a very simplified example of a kind of representational system that might reconstruct the perceptual similarity of

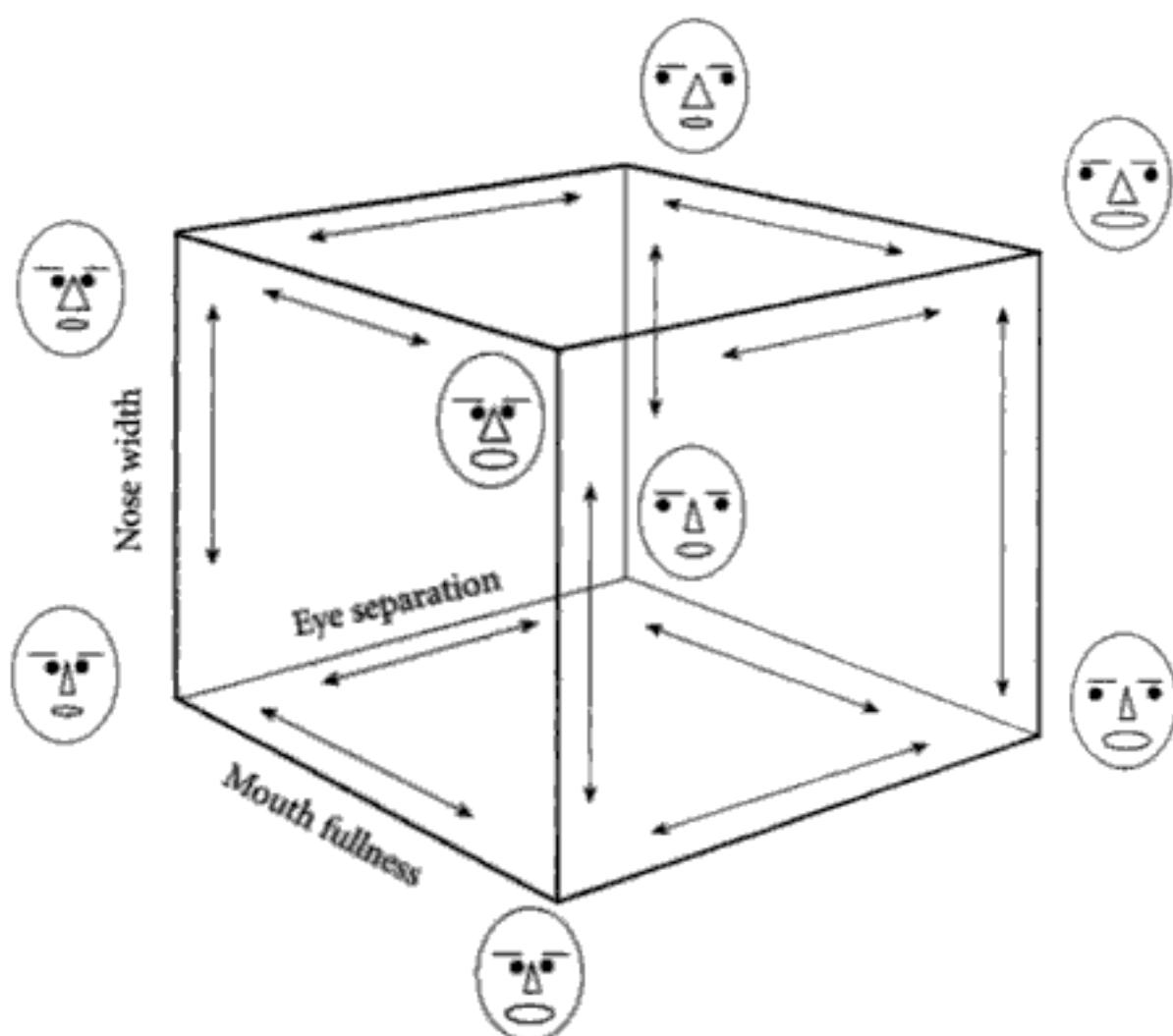


FIG. 9.1. A rudimentary face space of three dimensions. After Churchland (1995: 28, fig. 2.5); reproduced with permission of the author

faces. Roughly, there is an n -dimensional space in which each representation of a face occupies a proprietary region. The dimensions correspond to properties of the faces that are salient for the perception of similarity (and/or for similarity judgments). In the present case, the space is assumed to be three-dimensional for expository simplicity, 'but this is highly unrealistic—our coding space for faces probably has at least twenty dimensions—but it does evoke the wide range of faces one can discriminately code with only a few resources' (Churchland 1995: 29). The intended interpretation of the notation is, in effect, that locating a given face F at a given position P on a given dimension D corresponds exactly to representing F as having the property D to the extent P . The empirical claim is that, given the right choice of dimensions, the propinquity of the items in a space predicts their perceived or judged similarity.⁹

Several quick points about this kind of representation:

1. It doesn't matter that, in the example, a region of the space is identified with a representation(/concept) of a particular face. Any representation, either of an individual or of a property, would do equally well, so long as there is a natural and 'psychologically real' choice of salient dimensions along which to estimate the relative similarity of the things it represents.

2. In fact, to claim that a concept can be properly represented as a region in a vector space really amounts to *no more* than claiming that there does exist a natural and motivated choice of such dimensions. Looked at this way, the differences between vector-space encoding and more familiar accounts of meaning is easy to exaggerate. For example, a vector-space encoding is entirely compatible with the traditional doctrine that concepts are 'bundles' of semantic features. Indeed, the latter is a special case of the former, the difference being that, whereas semantic features are generally thought of as having

⁹ Churchland apparently thinks that the idea of developing a geometrical system of representation in which concepts, contents, and the like are identified with regions in a multidimensional semantic space is novel and indigenous to connectionist models of mind. '[It] contrasts sharply with the kinds of representational and processing strategies that . . . cognitive psychologists . . . have traditionally ascribed to us' (1989a: 171). In fact, semantic spaces have been around for quite a long while; see, e.g., Osgood *et al.* (1967). Most cognitive psychologists who worked with them eventually concluded that there isn't any question to which they are the answer. We think they were right to conclude that.

binary values, semantic dimensions are allowed to be continuous. Unsurprisingly, in light of this, semantic feature theories yield putative metrics of content similarity just as vector-space theories do: namely, the relative similarity of two concepts corresponds to a (possibly weighted) sum of the features they have in common.

3. The plausibility of identifying propinquities in a space with similarities of content depends entirely on the choice of the dimensions. Thus if you change one of the dimensions in Figure 9.1 so that what it expresses is not (for example) the width of a face's nose but the weight of the face's owner, the propinquities in that space will not predict judgments of perceptual similarity of the faces. This is the avatar, in the geometrical notation, of the perfectly general consideration that estimates of similarity must always be relativized to 'respects'.

So much, then, for the system Churchland wants to employ for annotating content relations among representations. Our main claim, in earlier phases of the exchange with him (see Churchland 1989b: 77–110; 1993: 667–72; 1996: 278–83; 1998: 5–32; Fodor and Lepore 1992, ch. 6; 1993d: 679–83), was that moving to this notation requires rephrasing a number of traditional cruxes that arise for identity-based theories of meaning, but that none of them is either resolved or bypassed. We briefly recapitulate the objections we raised.

As we have just seen, similarity of content is specified only with respect to a choice of the interpretation of the dimensions of a semantic space. There are three questions to be raised about such choices, all of which are patently counterparts to classical worries about individuating content in identity-based meaning theories.

How is the Choice of Semantic Dimensions to be Constrained?

For example, what determines that we are to taxonomize faces by the width of the noses they contain, rather than (for example) noses by some relation among the faces that contain them? Or, more plausibly, why is it right (if it is) to taxonomize the apple prototype by reference to its location in a space of colors (prototypical apples are red) rather than the red prototype by reference to its location in a space

of fruits (paradigm red is apple-red). This is exactly parallel to the problem that arises in identity-based meaning theories about what constrains the choice of primitive features, as in: Why is it right (if it is) to define 'mother' of in terms of 'parent' instead of the other way around?

Notice that this question is just as urgent for vector theories as it is for feature theories. For neither offers an interpretation of similarity relations, or indeed, of any semantic relations, among the primitive features/dimensions. Rather, they purport to capture relations of similarity of concepts relative to the primitive features/dimensions (that is, similarity in respect of the properties that the primitive features/dimensions express). Hence, *ceteris paribus*, the bigger the inventory of primitive features/dimensions the theories posit, the more semantic facts they fail to represent. In the limiting case, every concept corresponds to a primitive feature; and the theory becomes, in effect, disquotational. This sort of point has not, of course, gone unremarked in discussions of theories of meaning. The old-speak way to put it was: the notion of a semantic feature analysis of the lexicon is trivialized if there is allowed to be a distinct feature for each lexical item (so that the semantic analysis for the concept DOG says only that it is the unique concept that has the value *plus* on the feature dog). Nothing about this changes when you go over from binary features to continuous ones.

We suggested, earlier in the exchange with Churchland, that the reason this sort of problem seems not to bother him is that, like so many cognitive scientists, he believes deep down that only sensory properties can be semantic dimensions (Fodor and Lepore 1992). This does solve the problem, but at the cost of a recidivist and deeply implausible empiricism. If Churchland has now got a better way out, we cannot find it in his text.

What are the Truth Makers for Claims about the Semantic Interpretation of the Dimensions?

A second problem we raised for Churchland (Fodor and Lepore 1993d) was that the only metaphysics of meaning he seemed to have on offer was one according to which the inferential roles of representations are the truth makers for claims about their semantic

properties; and that could not be the right story about the semantic properties of the dimensions themselves.

As we saw above, to locate a concept (or a prototype) *C* with respect to a dimension *D* of a space is, in effect, to claim that the inference from being *C* to being *D* is constitutive (statistically or definitionally) of *C*. But if you thus identify the content of *C* by reference to its inferential relation to *D*, you cannot also identify the content of *D* by reference to its inferential relation to *C*. (You could, to be sure, identify the content of *D* with its position in some *higher-order* semantic space; but that would initiate a regress.) This is, once again, exactly parallel to a classic worry about identity-based meaning theories: if the contents of complex concepts are identified by their inferential relations to primitive concepts, the content of the primitive concepts has to be identified in *some other way*; and nobody knows what other way that is.

As far as we can tell, Churchland is now clear that the notion that contents are somehow individuated by their inferential roles will not work for the content of the dimensions. Rather, he thinks, the semantics of the dimensions is somehow to be reconstructed in terms of causal relations between, on the one hand, the representations that have values on the dimensions and, on the other, facts about the world. We have no objection to his taking this way out; but he does now need some reason for not treating such causal relations as the truth makers for *all* claims about the semantic properties of representations, thereby giving up, *inter alia*, the identification of contents with positions in semantic spaces. If (for example) the redness dimension has the content it does in virtue of causal relations between red things and the mental representation RED, why should the mental representation DOG not have the content it does in virtue of its causal relation to dogs? Note the exact correspondence to a traditional question about identity-based meaning theories of the 'two factor' variety: If it is to be assumed that a one-factor, causal theory works for the semantics of the primitive expressions, why is it to be denied that a one-factor causal theory will work for the semantics of all the expressions?

Strictly speaking, the vector theory identifies similarity of content only with respect to concepts that have rankings on the same dimensions. But it is implausible that this is true of the concepts people

actually entertain. This raises a problem for similarity-based meaning theories, which is exactly parallel to the traditional problem that identity-based theories have about how to avoid relativizing the individuation of concepts to the belief systems in which the concepts are embedded.¹⁰

Suppose, for example, your concept NEBRASKA is just like mine, except in the following respect: you think that Nebraska is about a thousand miles from Guadalajara; whereas I, never having heard of Guadalajara, have no views about the distance between Nebraska and it. So, the situation is that your semantic space, unlike mine, has a dimension *proximity to Guadalajara* along which the concept NEBRASKA has a location. Question: How much (if any) of this kind of thing can be tolerated, consonant with our concepts of NEBRASKA being similar? We emphasize that the vector notation as such offers no answer to this question. Transdimensional *similarity* (similarity among concepts in spaces which differ in dimensionality) is undefined; just exactly as (lacking an analytic/synthetic distinction) an inferential role semantics offers no interpretation of concept *identity* for minds that do not share all of their beliefs.

Likewise, some dimensions presumably count more than others do in similarity evaluations, and the relative weight of the dimensions differs for different concepts. So, at least in principle, it is possible for minds to differ either in respect of their weightings or in respect of the interaction between their weightings and the concepts, or both. Although you and I agree that literacy is a salient dimension for evaluating philosophers and engineers, I think it is important for philosophers to be well read, but it doesn't matter much whether engineers are; whereas you think it goes the other way around. Is there a principled answer to the question how (dis)similar this makes our concepts of engineers and philosophers? If so, what is the relevant principle?

There is, in short, a plethora of problems about how the

¹⁰ Maybe you don't mind if concepts are relativized to the belief systems in which they are embedded? So be it. But, of course, the semantic relativization problem lives just across the street from the semantic holism problem: If you have to relativize the identification of concepts to minds that have *some* of their beliefs in common, how do you avoid having to relativize the identification of concepts to minds that have *all* of their beliefs in common? Our point is that, if this worry is real, it persists unaltered when you change from a traditional semantics to a geometrical semantics.

similarity of concepts is to be evaluated across semantic spaces whose dimensionalities differ in various ways. These are just the old problems about the relativization of concepts to belief systems; except that, in the new notation, talk about the way that a concept is positioned with respect to the dimensions of a semantic space replaces talk of the way that a concept is embedded in a system of beliefs. Whichever notation you prefer, you have the same problems about how to identify concepts and the like in the sorts of circumstances that typically hold *across minds* (or 'languages', or 'theories').

We suspect that such problems are unsolvable unless the analytic/synthetic problem can be solved. Nonetheless, as far as we can tell, this is the issue to which Churchland takes his new theory to be germane. We are going to argue that this is a massive *ignoratio elenchi* on Churchland's part. The proposal he has on offer is not of *the right logical type* to be a solution to the problem of transdimensional content individuation; and, though there is a problem in the general vicinity that it is of the right logical type to solve, it doesn't, in fact, solve that problem either.

Churchland's proposal co-opts work by Aarre Laakso and Gary Cottrell (1998), which, he says, 'constitute[s] a decisive answer to Fodor and Lepore's challenge . . . Laakso and Cottrell . . . successfully deploy one member of a large family of mathematical measures of conceptual similarity, measures that see past differences—even extensive differences . . .' (Churchland and Churchland 1998: 81; Churchland 1998: 5), and now, given the nature of the challenge that we had offered (namely, that there is a relativization problem for vector-space semantics just like the relativization problem for identity-based semantics), one expects this sentence to end 'of the *semantic* dimensions of the conceptual spaces'. For, to repeat, a vector-space semantics identifies concepts by reference to their positions in a semantic space; and the problem we raised—the one that Churchland apparently thinks Laakso and Cottrell have solved—was precisely 'How can minds that do not have the same semantic spaces have the same (or similar) concepts?' Here, however, is the way the sentence that we just quoted actually does end: 'measures that see past differences . . . in the connectivity, the sensory inputs and the *neural* dimensionality of the networks being compared' (ibid.; our emphasis).

But something has gone wrong here. Neural spaces and semantic

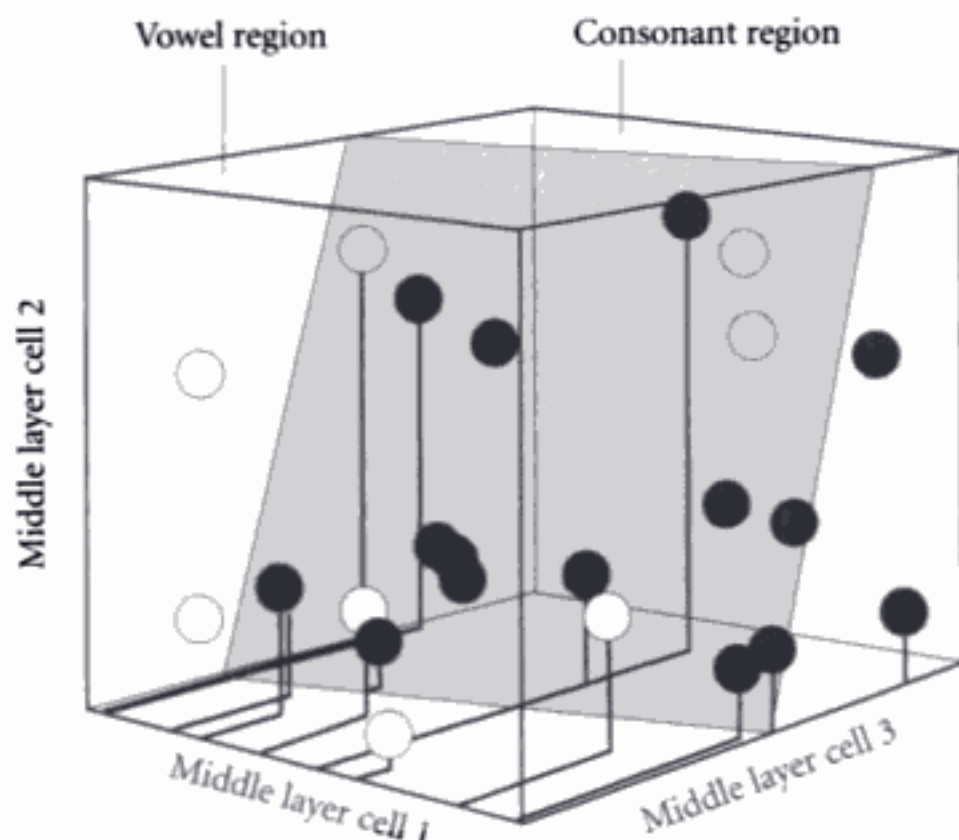


FIG. 9.2. The activation space of the cells at the middle layer of NETtalk. From Churchland (1995: 88, fig. 4.22); reproduced with permission of the author

spaces are quite different sorts of things. As we have been seeing, positions along the dimensions of the semantic space relative to which a concept is situated correspond to propositional attitudes the holding of which is supposed to be constitutive of having the concept. By contrast, the dimensions in a neural space *correspond to neurons*; and the positions along a given dimension specify the degree of excitation of the corresponding neuron (at a time).¹¹ Unsurprisingly, though he is not reliable in observing it, this difference does show up in Churchland's notation now and then; see Figure 9.2. Notice that the axes of *this* cube are labeled with the names of neurons, not with the names of semantic dimensions.

Now, just as there is a relativization problem about the identity (or similarity) of *concepts* and the like across *semantic* spaces of different

¹¹ If we simplify (as, indeed, connectionist models often do) by assuming binary dimensions as opposed to ones with real number values, a state description relative to a neural space is an assignment of 1 or 0 to each neuron depending on whether it is active (at *t*).

dimensionality, so, too, there is a relativization problem about the identity (or similarity) of *brain states* across *neural spaces* of different dimensionality. For example, if you are a connectionist, you identify circuits with *n*-tuples of nodes and links; and the nodes and links in a network are themselves identified *by the totality of their relations to one another*. On this picture, the notion 'same(/similar) circuit' is defined *only* for networks of identical dimensionality, just like the notion 'same(/similar) content'.

We think that this analogy between individuation problems has caused Churchland to confuse the question of how to provide a transdimensional notion of brain-state identity with the question of how to provide a transdimensional notion of content identity: it is because he has failed to observe this distinction that he seems indifferent whether the dimensions of the spaces he talks about are semantic or neural. It is likewise because of this confusion, and because Churchland thinks (wrongly; see below) that Laakso and Cottrell have solved the transdimensional individuation problem for brain states that he believes (again wrongly) that Laakso and Cottrell have solved the transdimensional individuation problem for content states. So we shall now argue.

Let's do this slowly. There is, to repeat, a problem about type-individuating brain states across neural spaces of different dimensionality,¹² just as there is a problem about individuating contents

¹² For that matter, there is a problem about typing brain states across neural spaces with the *same* dimensionality if the dimensions of the spaces are supposed to correspond to neurons as *anatomically* individuated. There really is, we think, a dilemma hereabouts. If a neuron is a *physiological* unit (for example, neurons are individuated by their connectivity to other neurons and the like), then there is a holism problem for them just as there is for blatantly functional entities like circuits. On the other hand, if neurons are *anatomical* units, there is no guarantee that brains that have the same neural populations are ever in the same brain states by any functional criterion. As far as we can tell (for example, from the labeling of Fig. 9.2, his Fig. 4.22), Churchland prefers the first horn of this dilemma; that is, he (implicitly) takes the individuation of neurons to be sensitive to their functions. This raises, but does not solve, the question how exactly the dimension labels of his neural activation spaces are to be interpreted; that is, *which* functional properties 'being neuron A' and the like express.

Begging the question whether, or to what extent, the 'neurons' in their networks are supposed to be *anatomical* as well as *functional* units is a standard practice in connectionist texts. It makes it seem to uninitiates that connectionism is brain science, not just a species of computational psychology. That is very good for getting grants, but it does cause a lot of confusion.

across semantic spaces of different dimensionality. But it cannot simply be taken for granted that a proposed solution for the first of these problems would solve (or even help to solve) the second. Suppose you have some scheme for picking out homologous brain regions across neural spaces of different dimensionality and that you are convinced that it preserves similarity (or identity) of neural types. That is, homology is defined so that if there is a region of an n -dimensional space such that:

1. All the brain states in this region are neurologically similar or identical; and if
2. for every space of m dimensions ($n < m$) there is a homologous region such that, if a brain state has a location in the n -space it has a location in the homologous region of the m -space; then:
3. the brain states in the homologous regions of m -space are likewise neurologically similar or identical.

That is, of course, a very strong condition on a notion of trans-dimensional homology for regions of neurological spaces. But our point is that, even if you had a notion of transdimensional homology which, in this sense, preserved *neural-state* similarity (identity), you would have no right to assume that it likewise preserves *semantic* similarity (or identity). If you want to claim that it does, you will need to provide an argument. Roughly, you need to show why, if the brain states in a region of a certain neurological space are similar or identical *in content*, then so, too, are the brain states in a region of any higher-dimension neurological space that is homologous under the preferred transdimensional mapping.¹³

¹³ It is worth noting that the claim that content-state types correspond to brain-state types under some or other criterion of individuation of the latter is implausibly strong quite aside from issues of *transdimensional* individuation. In particular, such a claim would require that wherever states of mind overlap in their intentional contents, there is also an overlap of the corresponding brain states. So, for example, there would have to be a (constituent) brain state common not just to tokens of the thought *that is red*, but also to tokens of the thought *I have decided to paint the cat red*; and to tokens of the thought *red is as far from green as blue is from yellow*; and so forth. There is, as far as anybody knows, not the slightest reason for supposing that the relation between mental contents and brain states meets this constraint even if issues of transdimensionality are ignored. In fact, as things now stand, there are *no* known examples of cases where it does. It is a nice irony that it fails notoriously to be satisfied by the relation between circuits in connectionist networks and the contents that the circuits are supposed to encode. (That is exactly why connectionists have so much trouble with compositionality.)

Notice that taking physicalism for granted (which, for the sake of the argument, we are happy to do) does *not* provide the argument that is required. 'Token' physicalism says that, if individual brain states are identical, then so, too, are the corresponding individual intentional states; but token physicalism says nothing about whether there are such things as transdimensional counterparts of brain-state tokens; or whether, if there are, they *ipso facto* share any of their semantic properties; or, if they do, which ones. Likewise, 'type' physicalism guarantees that, if brain states are of the very same kind, then so too are the corresponding content states. But it makes no commitments about what, if any, content relations obtain for brain states that are merely *transdimensionally homologous* according to some or other typological scheme. Churchland must be assuming that there is a typology of brain states that meets the condition that when neurological similarity (or identity) is preserved under a transdimensional mapping, content identity is preserved, too. Nothing else would warrant his persistent practice of putting *semantic* labels on the dimensions of *neural* spaces and vice versa. But, to repeat, that semantic similarity is preserved transdimensionally when neural similarity is preserved transdimensionally does not follow from any kind of physicalism we have heard of, nor would it seem to be remotely plausible on independent grounds.

Here, to summarize, is the dialectical situation as we understand it so far: Churchland thinks that Laakso and Cottrell have provided a notion of 'kind of brain state' which preserves similarity of neural-state types across spaces which differ in their neural dimensionality. And he (tacitly) infers that they have *thereby* provided a transdimensional notion of content-state similarity as well. But, to put it mildly, this inference is tendentious. Churchland needs an argument for drawing it, but provides none. We doubt that he has one that he is shy about producing. Rather, our guess is that he has simply confused *semantic* spaces (of which the typical occupants are intentional states, so described) with *neural* spaces (of which the typical occupants are brain states, so described).

All right so far. We now propose to show that, in fact, the sort of criterion Laakso and Cottrell suggest for type-individuating brain states across dimensionality differences pretty clearly does *not* preserve either identity or similarity of the contents of the

states.¹⁴ Basically, Churchland's idea is that, even if my neural space has fewer dimensions than yours, we can embed the one in the other so long as the dimensions of my space are a subset of the dimensions of yours. In which case, we can assign each of my (actual or possible) neural states to the same state type as some (actual or possible) neural state of yours. *A fortiori*, we can type identify (at least some) brain states across (at least some) differences among neurological spaces. That, as far as we can tell, is the burden of the following remark:

If a pair of dimensionally diverse spaces both contain . . . a same-shape n -dimensional hypersolid . . . then both spaces must each either constitute or contain an n -D subspace, an n -D hyperplane of some orientation or other, that confines each hypersolid at issue. Since those respective hyperplanes . . . have the same dimensionality, there will be no problem in comparing the two solids. (Churchland and Churchland 1998: 91; Churchland 1998: 13–14)

We paraphrase (under correction, to be sure):

Claim Q. If my neural space is properly embedded in yours, then (*ceteris paribus*) my (actual or possible) brain states are a subset of yours.

Suppose, for the moment, that Q is true; hence that, if a neural space of lower dimensionality is embedded in a neural space of a higher dimensionality, then each brain state that has a location in the former space has a corresponding location in the latter. Our argument will be that, even if Q is true, it *does not provide for the transdimensional individuation of content states*. That is because R—which is the analog of Q for *semantic* spaces—is clearly false; content identity is *not* preserved under the embedding of one semantic space in another:

Claim R. If my semantic space is embedded in yours, then (*ceteris paribus*) my (actual or possible) concepts are a subset of yours.

¹⁴ *A fortiori*, if the Laakso and Cottrell sort of view is right about how to individuate brain-state types, then type physicalism is false about the relation between brain-state types and intentional-state types. (Type physicalists should not be alarmed by this, however; we shall see presently that the proposal on offer for transdimensionally individuating brain-state types doesn't work either.)

Here is a candidate counterexample to R (the Nebraska/Guadalajara case discussed above is another). Suppose I have only three semantic dimensions, hard-soft, black-white, heavy-light. And suppose my concept ROCK is identified with a vector that specifies a region in the space that these dimensions define. Likewise for you: your semantic space also has these dimensions (*inter alia*); and you too have a concept that lies within the boundaries of the space that they define. The difference between us is that, whereas these dimensions define *the whole* of my semantic space, they define only an embedded subspace of yours. Now, does it follow that our concepts ROCK are similar? We shouldn't have thought so. For perhaps your space has a dimension ANIMACY which, by assumption, mine lacks. And suppose that you think that rocks are actually quite animate; much more animate, say, than turtles; at least as animate, indeed, as half the members of the Daughters of the American Revolution. Is it *still* the case that your ROCK concept is similar to mine? If so, suppose also that your space contains a dimension for abstractness, and that you think that rocks are pretty abstract; at least as abstract as the natural numbers, say, though maybe less abstract than the real numbers. Is your concept ROCK *still* similar to mine? If there are *any* such cases where the right answer is 'no', Churchland loses.

Our point is that, even if concepts, semantic properties, intentional states, and the like can be identified with positions in a state space, and even if position in that state space is preserved when the latter is embedded in a space of higher dimension, it does not follow that those (or indeed any) semantic properties, intentional states, and so on can be identified with positions in the higher-order space. That is because the conditions for the individuation of contents are non-monotonic. This, too, is something that everybody already knew; indeed, it is just the analytic/synthetic problem in one of its familiar guises. Presumably Homer has a concept of water which was similar to ours as far as it went.¹⁵ (Water is wet, it is potable, it freezes in the winter . . . and so forth.) There remains the question whether Homer's concept of water was similar to ours *tout court*, and the relevant considerations are much the same ones that the corresponding

¹⁵ Assuming, as Churchland of course does, that what concepts you have depends on what beliefs you have.

question about the *identity* of our concepts would turn on: our WATER concept is embedded in a theory (a *Weltanschauung*; a form of life; blah, blah, blah) many of whose dimensions are quite different from any of the ones in Homer's semantic space. What, if anything, follows about which concepts we share with him?

Recapitulation: even if we did have a transdimensional notion of 'brain-state similarity', it would not follow that we would then have a transdimensional notion of 'content similarity'. That is, even if regions of a lower dimensional space can be securely correlated with semantic interpretations, there is no reason why such a correlation should survive any proposed scheme of transdimensional mapping.

Conclusions

We close by remarking that, putting questions about content to one side, Churchland's sort of story really does not work even for the transdimensional typing of brain states. Q is false because the identity conditions for neural-state types, like the identity conditions for content-state types, is nonmonotonic. A brain state that is identified with a position in a neural space of m dimensions need not be type-identical with any region in an n -dimensional neural space in which the m -space is embedded. This is because brain-state types are functionally individuated; and functional individuation is itself nonmonotonic.¹⁶

Suppose we have a neural space of m dimensions; and suppose that the function of a certain activation state of these neurons is to cause a flight response. Now, embed the m -space in one of higher neural dimensionality (which is to say: put the same m -neurons in a bigger brain, preserving their connectivity to one another). It surely does not follow that there is any activation state of these (or, indeed, of any) neurons in the larger brain whose function is to cause flight responses. Indeed, it does not even follow that activation states defined over the original m -neurons have *any function at all* in the

¹⁶ We are not, please note, asserting—or even suggesting—that content individuation is nonmonotonic *because* it is a species of functional individuation. Heaven forbid.

larger brain; that depends, *inter alia*, on what other neurons they are connected to. It is, for example, perfectly possible there is no physiologically distinguishable system that all or only the original *m*-neurons belong to when they are in the bigger brain.¹⁷

Brief conclusion: the question we were worried about in our earlier exchanges with Churchland was: 'How can there be *content similarities* between minds that have different beliefs?' We were worried about this because Churchland keeps saying that similarity semantics avoids the main problems with which analytic/synthetic semantics is plagued, and our understanding is that 'How can there be content identities between minds that have different beliefs?' is the main problem with which analytic/synthetic semantics is plagued.

But the question Churchland actually offers an answer to is: 'How can creatures be in similar brain states if they have different numbers of neurons?' This is an interesting kind of question, to be sure; but it has nothing in particular to do with meaning. To the contrary, it arises for any theory that is committed to the functional individuation of things over which it quantifies. This is so whether or not the things over which it quantifies are psychological/intentional/semantic. In particular, the question of transdimensional functional-state individuation arises for any theory that is committed to kinds of states that can have multiple realizations. Compare 'How can tubes and transistors both be amplifiers?', 'How can jets and diesels both be engines?', 'How can a wood thing and a plastic thing both be pawns?', and so on.

It would be nice to have a transdimensional, similarity-preserving, type-identity criterion for brain states, and it would also be nice to have a transdimensional, similarity-preserving, type-identity criterion for content states. And it would be very nice (and very surpris-

¹⁷ Suppose, however, that it is denied that the individuation of kinds of brain states is functional. Even so, it is far from obvious that either identity or similarity of brain-state kinds would be preserved under the sort of transdimensional mapping Churchland imagines. It is patently not the case that brain states that are similar *in a given world* because they share certain nonfunctional properties are *ipso facto* similar in *any world* in which they have those properties in common. There may be ways of choosing canonical neural dimensions that meet this condition; but there is no *a priori* reason to believe that there are. To put this another way: it is one question whether brain states are individuated by their function; it is quite a different question whether, however they are individuated, they are reducible to neural states.

ing) if the two criteria turned out to be the same. But, as far as anybody knows or has any reason to believe, these two individuation problems raise quite different ontological issues, and their solutions—if, indeed, they have got solutions—are likely to be largely independent. All they have in common, as far as we can tell, is that both kinds of individuation are nonmonotonic, and that Churchland hasn't got a workable account of either. And nor, of course, has anybody else.

Appendix 9.1 Afterthought

A recent note (Abbott 2001) catches a mismatch between our formulation of (S3) and the illustrative example of it that we provided. We are grateful to Abbott, but we don't think the mistake is serious, and we stand by the claim made earlier in this chapter: If you ground a theory of language in meaning similarity, rather than meaning identity, then you must abandon the idea that lexical meanings are context-independent and (hence) the idea that the meanings of complex linguistic expressions are compositional.

Abbott is right that the example we give in the earlier section 'Compositionality' is not a counterexample to (S3): *If m is similar to part of the meaning of 'a' and 'a' is a constituent of 'b', then m is similar to part of the meaning of 'b'*. Nevertheless, (S3) is false. Here's a better example: part of the meaning of 'explosion' ($= m$) is similar to part of the meaning of 'leopard' ($= a$) since both contain *dangerous*. 'Leopard' is a constituent of 'dead leopard' ('tame leopard', etc.; $= b$). But no part of the meaning of 'explosion' is similar to any part of the meaning of 'dead leopard'; dead leopards aren't dangerous.

It bears emphasis that whether (S3) is true depends very much on how you read 'part of the meaning'. If you read it as *overlap of entailment*, then (S3) is true. *Feline* is, let's say, entailed by 'leopard'; 'leopard' is a constituent of 'dead leopard' and, since dead leopards are dead felines, *feline* is likewise entailed by 'dead leopard'. But it is essential to seeing what is going on in the argument with Churchland that he does *not* explicate meaning similarity in terms overlapping entailments. Rather, he understands it as propinquity of stereotypes in a semantic space. That 'leopard' and 'explosion' have nearby positions on the *dangerous* vector is just the sort of case that he has in mind.

The relevant consideration is that overlap of entailment is a notion that presupposes meaning *identity* (not just meaning similarity): P1 and P2 overlap in entailments just in case some entailment of P1 is the same proposition as (is identical in meaning to) some entailment of P2. But it is the whole point of Churchland's undertaking to provide a notion of meaning

similarity that does *not* presuppose a notion of meaning identity; and thereby to save meaning from the familiar Quinean strictures. Likewise with other philosophers and psychologists who are interested in prototypes, stereotypes, and the like as models for concepts. But to repeat: when so understood, similarity in meaning to a constituent does not guarantee similarity in meaning to expressions that contain it.

REFERENCES

- ABBOTT, B. (2000), 'Fodor and Lepore on Meaning, Similarity and Compositionality', *Journal of Philosophy*, 97/8 (Aug.), 454–5.
- BENCIVENGA, E. (1980a), 'Truth, Correspondence, and Nondenoting Singular Terms', *Philosophia*, 219–29.
- (1980b), 'Free Semantics for Definite Descriptions', *Logique et Analyse*, 23: 393–405.
- (1991), 'Free Semantics', in K. Lambert (ed.), *Philosophical Applications of Free Logic* (New York: Oxford University Press), 98–110.
- BLOCK, N. (1986), 'Advertisement for a Semantics for Psychology', in P. French, T. Uehling, and H. Wettstein (eds.), *Midwest Studies in Philosophy*, vol. 10: *Studies in the Philosophy of Mind* (Minneapolis: Minnesota University Press), 615–78.
- BRANDON, R. (2000), *Articulating Reasons* (Cambridge, Mass.: Harvard University Press).
- CHOMSKY, N. (1965), *Aspects of the Theory of Syntax* (Cambridge, Mass.: MIT Press).
- CHURCHLAND, P. M. (1979), *Scientific Realism and the Plasticity of Mind* (Cambridge: Cambridge University Press).
- (1989a), *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science* (Cambridge, Mass.: MIT Press).
- (1989b), 'Some Reductive Strategies In Cognitive Neurobiology', in Churchland (1989a: 77–110).
- (1989c), 'Perceptual Plasticity and Theoretical Neutrality: A Reply to Jerry Fodor', in Churchland (1989a: 255–80).
- (1993), 'Fodor and Lepore: State-Space Semantics and Meaning Holism', *Philosophy and Phenomenological Research*, 53: 667–72.
- (1995), *The Engine of Reason: The Seat of the Soul* (Cambridge, Mass.: MIT Press).
- (1996), 'Second Reply to Fodor and Lepore', in R. McCauley (ed.), *The Churchlands and their Critics* (Cambridge: Blackwell), 278–83.
- (1998), 'Conceptual Similarity across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered', *Journal of Philosophy*, 95/1: 5–32.
- and CHURCHLAND, P. S. (1998), *On the Contrary: Critical Essays, 1987–1997* (Cambridge, Mass.: MIT Press).

- CRUSE, D. A. (1986), *Lexical Semantics* (Cambridge: Cambridge University Press).
- DAVIDSON, D. (1984), 'Truth and Meaning', in his *Essays on Truth and Interpretation* (Oxford: Oxford University Press), 17–42.
- DRETSKE, F. (1981), *Knowledge and the Flow of Information* (Cambridge, Mass.: MIT Press).
- DUMMETT, M. (1975), 'Wang's Paradox', *Synthese*, 30: 301–24.
- (1978), *Truth and Other Enigmas* (Cambridge, Mass: Harvard University Press), 248–68.
- (1993), 'What is a Theory of Meaning?', in his *The Seas of Language* (Oxford: Oxford University Press), 1–94.
- ELMAN, J., BATES, E., JOHNSON, M., and KARMILOFF-SMITH, A. (1996), *Rethinking Innateness: A Connectionist Perspective on Development* (Cambridge, Mass.: MIT Press).
- FIELD, H. (1977), 'Logic, Meaning, and Conceptual Role', *Journal of Philosophy*, 74: 379–409.
- FINE, K. (1975), 'Vagueness, Truth and Logic', *Synthese*, 30: 265–300.
- FODOR, J. A. (1970), 'Three Reasons for not Deriving "Kill" from "Cause to Die"', *Linguistic Inquiry*, 1: 429–38.
- (1981), 'The Present Status of the Innateness Controversy', in his *Representations* (Cambridge, Mass.: MIT Press), 257–316.
- (1990), *A Theory of Content and Other Essays* (Cambridge, Mass.: MIT Press).
- (1994), *The Elm and the Expert* (Cambridge, Mass.: MIT Press).
- (1998a), *Concepts* (Oxford: Oxford University Press).
- (1998b), *In Critical Condition* (Cambridge, Mass.: MIT Press).
- and LEPORE, E. (1992), *Holism, a Shopper's Guide* (Oxford: Blackwell).
- (1993a), 'The Normativity of the Intentional', in B. Dahlbom (ed.), *Dennett and his Critics* (Oxford: Blackwell), 70–82.
- (1993b), 'Replies to Brandom, Churchland, Devitt, McLaughlin, and Rey on Holism', *Philosophy and Phenomenological Research*, 53 (Sept.), 673–83.
- (1993c), 'Replies to Block and Boghossian', *Mind and Language*, 8: 41–8.
- (1993d), 'Reply to Churchland', *Philosophy and Phenomenological Research*, 53: 679–83.
- (1994a), 'Is Radical Interpretation Possible?', *Philosophical Perspectives*, 8: 101–19.
- (1994b), 'Meaning, Holism and the Problem of Extensionality', in G. Preyer (ed.), *Language, Mind and Epistemology* (Deventer: Kluwer), 3–19.

- (1994c), 'Replies to 11 Critics on *Holism*', *Grazer Philosophica*, 46: 303–22.
- (1994d), 'Replies to Perry and Boghossian on *Holism*', *Philosophical Studies*, 77: 139–47.
- (1994e), 'What is the Connection Principle?', *Philosophy and Phenomenological Research* (Dec.), 837–46.
- (1996a), 'Paul Churchland and State Space Semantics', in R. McCauley (ed.), *The Churchlands and their Critics* (Oxford: Blackwell), 145–58.
- (1996b), 'The Pet Fish and the Red Herring: Why Concepts aren't Prototypes', *Cognition*, 58/2 (Feb.), 243–76.
- (1996c), 'What Can't Be Valued Can't Be Valued, and it Can't Be Supervalued Either', *Journal of Philosophy*, 93: 516–36.
- (1998), 'The Emptiness of the Lexicon: Reflections on Pustejovsky', *Linguistic Inquiry*, 29/2 (Spring), 269–88.
- (1999a), 'All at Sea in Semantic Space: Churchland on Meaning Similarity', *Journal of Philosophy*, 96/8 (Aug.), 381–403.
- (1999b), 'Impossible Words?' *Linguistic Inquiry*, 30/3 (Summer), 445–53.
- (2001), 'Brandom's Burdens: Compositionality and Inferentialism', *Philosophy and Phenomenological Research*, 73/2 (Sept.), 465–81.
- and PYLYSHYN, Z. (1988), 'Connectionism and Cognitive Architecture: A Critical Analysis', *Cognition*, 28: 3–71.
- GOPNICK, A., and MELTZOFF, A. N. (1997), *Words, Thoughts and Theories* (Cambridge, Mass.: MIT Press).
- HALE, K., and KEYSER, S. J. (1993), 'On Argument Structure and the Lexical Expression of Syntactic Relations', in K. Hale and S. J. Keyser (eds.), *The View from Building 20: Essays on Linguistics in Honor of Sylvain Bromberger* (Cambridge, Mass.: MIT Press), 53–109.
- HARMAN, G. (1973), *Thought* (Princeton: Princeton University Press).
- HIGGINBOTHAM, J. (1994), 'Priorities of Thought', *Supplementary Proceedings of the Aristotelian Society*, 68: 85–106.
- HORWICH, P. (1998), *Meaning* (New York: Oxford University Press).
- HUTTENLOCHER, J., and HEDGES, L. (1994), 'Combining Graded Categories', *Psychological Review*, 101/1: 157–65.
- JACKENDOFF, R. (1988), 'Conceptual Semantics', in U. Eco, M. Santambroggio, and P. Violi (eds.), *Meaning and Meaning Representations* (Bloomington: Indiana University Press), 81–97.
- KAMP, H. (1975), 'Two Theories about Adjectives', in E. L. Keenan (ed.), *Formal Semantics of Natural Language* (New York: Cambridge University Press), 123–55.

- KAMP, H. (1981), 'The Paradox of the Heap', in U. Monnich (ed.), *Aspects of Philosophical Logic* (Boston: Reidel), 225–77.
- and PARTEE B. (1995), 'Prototype Theory and Compositionality', *Cognition*, 57: 129–91.
- KARMILOFF-SMITH, A. (1992), *Beyond Modularity: A Developmental Perspective on Cognitive Science* (Cambridge, Mass.: MIT Press).
- KATZ, J., and FODOR, J. A. (1963), 'The Structure of a Semantic Theory', *Language*, 39: 170–210.
- KRIPKE, S. (1980), *Naming and Necessity* (Cambridge, Mass.: Harvard University Press).
- LAASKO, A., and COTTRELL, G. (1998), 'Qualia and Cluster Analysis: Assessing Representational Similarity between Neural Systems', MS.
- LEPORE, E. (1994), 'Conceptual Role Semantics', in S. Guttenplan (ed.), *A Companion to the Philosophy of Mind* (Oxford: Blackwell), 192–8.
- (1996), 'Conditions on Understanding Language', *Proceedings of the Aristotelian Society*, 97/1: 41–60.
- LEWIS, D. (1972), 'General Semantics', in D. Davidson and G. Harman (eds.), *Semantics of Natural Language*, 2nd edn. (Boston, Mass.: Reidel), 169–218.
- MCGEE, V., and MCLAUGHLIN, B. (unpublished), 'The Value of Supervaluations', MS.
- MATES, BENSON (1950), 'Synonymity', *University of California Publications in Philosophy*, 25; repr. in L. Linsky (ed.), *Semantics and Philosophy of Language* (Urbana: University of Illinois Press, 1952), 111–38.
- MILNE, A. A. (1957), *The World of Pooh* (New York: E. P. Dutton).
- MORREAU, M. (1999), 'Supervaluation can Leave Truth-Value Gaps After All', *Journal of Philosophy*, 96: 148–56.
- OSGOOD, C. E., SUCI, G. J., and TANNENBAUM, P. H. (1967), *The Measurement of Meaning* (Urbana: University of Illinois Press).
- OSHERSON, D. N., and SMITH, E. E. (1981), 'On the Adequacy of Prototype Theory as a Theory of Concepts', *Cognition*, 9: 35–58.
- (1988), 'Conceptual Combination with Prototype Concepts', in A. Collins and E. Smith (eds.), *Readings in Cognitive Science* (San Mateo, Calif.: Morgan Kaufman), 323–35.
- PUSTEJOVSKY, J. (1995), *The Generative Lexicon* (Cambridge, Mass.: MIT Press).
- PUTNAM, H. (1975), 'The Meaning of "Meaning"', in his *Mind, Language, and Reality* (New York: Cambridge University Press), 215–71.
- ROLF, B. (1984), 'Sorites', *Synthese*, 56: 219–50.
- ROSCH, E. H. (1973), 'Natural Categories', *Cognitive Psychology*, 4: 328–50.
- RUSSELL, B. (1905), 'On Denoting', *Mind*, 14: 479–93.

- SAINSBURY, R. M. (1988), *Paradoxes* (New York: Cambridge University Press).
- SANFORD, D. (1976), 'Competing Semantics of Vagueness: Many Values versus Super-Truth', *Synthese*, 33: 195–210.
- SCHIFFER, S. (1987), *Remnants of Meaning* (Cambridge, Mass.: MIT University Press).
- SKINNER, B. F. (1957), *Verbal Behavior* (New York: Appleton Century Cross).
- SKYRMS, B. (1968), 'Supervaluations: Identity, Existence, and Individual Concepts', *Journal of Philosophy*, 69/16: 477–82.
- (1970), 'Return of the Liar: Three-Valued Logic and the Concept of Truth', *American Philosophical Quarterly*, 7: 153–61.
- SLOMAN, S., and RIPS, L. J. (1998), *Similarity and Symbols in Human Thinking* (Cambridge, Mass.: MIT Press).
- SMITH, E. E., and MEDIN, D. L. (1981), *Categories and Concepts* (Cambridge, Mass.: Harvard University Press).
- STICH, S. (1981), 'On the Ascription of Content', in A. Woodfield (ed.), *Thought and Object* (New York: Oxford University Press), 153–206.
- (1983), *From Folk Psychology to Cognitive Science* (Cambridge, Mass.: MIT Press).
- STRAWSON, P. F. (1950), 'On Referring', *Mind*, 59: 320–44.
- TAPPENDEN, J. (1993), 'The Liar and Sorites Paradoxes: Toward a Unified Treatment', *Journal of Philosophy*, 90/1: 551–77.
- VAN FRAASSEN, B. (1966a), 'Singular Terms, Truth-Value Gaps, and Free Logic', *Journal of Philosophy*, 63/17: 481–95.
- (1966b), 'The Completeness of Free Logic', *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 12: 219–34.
- (1968), 'Presuppositions, Implications, and Self-Reference', *Journal of Philosophy*, 65/5: 136–52.
- (1969), 'Presuppositions, Supervaluations, and Free Logic', in K. Lambert (ed.), *The Logical Way of Doing Things* (New Haven: Yale University Press), 70–1.
- and LAMBERT, K. (1967), 'On Free Description Theory', *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*, 13: 225–40.
- VON KUTSCHERA, F. (1975), 'Partial Interpretations', in E. Keenan (ed.), *Formal Semantics of Natural Language* (Cambridge, Mass.: Cambridge University Press), 156–74.
- WEINREICH, U. (1959), 'Travels through Semantic Space', *Word*, 14: 346–66.
- WILLIAMSON, T. (1994), *Vagueness* (New York: Routledge), 142–64.
- WRIGHT, C. (1987), 'Further Reflections of the Sorites', *Philosophical Topics*, 15: 227–90.
- ZIFF, P. (1960), *Semantic Analysis* (Ithaca: Cornell University Press).

NAME INDEX

- Abbott, B. 199
- Barwise, Jon 11
- Bencivenga, E. [82](#), [83](#)
- Berkeley, George 171, 172
- Block, Ned 174
- Brandom, Robert [28](#), [135](#), [135-54](#)
- Chihara, Charles [84](#)
- Chomsky, Noam [146](#), [155](#), [166](#)
- Churchland, Paul M. [155-73](#),
174-200
- Cottrell, G. 190, 192, 194, 195
- Craig, William [84](#)
- Cruse, D. A. [90](#)
- Davidson, Donald [153](#)
- Dennett, Daniel [150](#)
- Derrida, Jacques [23](#)
- Dummett, M. [45](#), [65](#), [68-70](#), [75](#), [84](#)
- Field, Hartry [17](#)
- Fine, Kit [64-70](#), [73-6](#), [84](#)
- Fodor, Jerry A. [11](#), [27](#), [43](#), [60](#), [90](#), [93](#),
[95](#), [101](#), [117](#), [139](#), [143](#), [175](#), [181](#),
186-7, 190
- Frege, G. [11-12](#), [137-9](#), [144](#), [168](#)
- Goble, Lou [84](#)
- Gopnik, A. [3](#)
- Hale, K. [120-30](#)
- Harman, Gilbert 174
- Hedges, L. [41](#), [42](#)
- Hegel, George W. F. [137-8](#), [154](#)
- Higginbotham, J. [99](#)
- Horwich, Paul [43-62](#)
- Hume, David [5](#), [172](#)
- Huttenlocher, J. [41](#), [42](#)
- Jackendoff, R. 174
- Johnson, Samuel [96](#)
- Kamp, H. [27](#), [29](#), [31-5](#), [37-8](#), [63](#), [65](#),
[67-70](#)
- Kant, Emanuel [5](#), [137-9](#)
- Karmiloff-Smith, Annette [3](#)
- Katz, J. [93](#), [117](#)
- Keyser, S. J. [120-30](#)
- Kuhn, Thomas [23](#)
- Laasko, A. 190, 192, 194-5
- Lambert, K. [79](#)
- Land, Edwin [159](#), 165
- Lepore, Ernie [90](#), [95](#), [139](#), 175, 181,
186-7, 190
- Lewis, D. [65](#), [84](#)
- Locke, John 171
- Mates, Benson [96](#)
- McGee, Vann [77](#), [84-5](#)
- McLaughlin, Brian [84-5](#)
- Medin, D. L. [27](#)
- Meltzoff, A. N. [3](#)
- Millikan, Ruth 11
- Milne, A. A. [101](#)
- Morreau, M. [84-6](#)
- Oliver, Alex 154
- Osgood, C. E. 185
- Osherson, D. N. [27](#), [30](#), [32](#), [35-7](#)
- Parsons, Terry [84](#)
- Partee, B. [27](#), [29](#), [31-5](#), [37-8](#), [63](#), [67](#)
- Penco, Carl 154
- Perry, J. 11
- Pustejovsky, J. [89-118](#)
- Putnam, Hillary [23](#), [34](#), [178](#)
- Pylyshyn, Zenon [143](#)
- Quine, W. V. O. [19-22](#), [24-5](#), [69](#), [94](#), [96](#),
[140](#), [156-9](#), 173, 200
- Rips, L. J. 174
- Rolf, B. [68](#)

- Rorty, R. [23](#)
Rosch, E. [H. 34](#)
Russell, Bertrand [78](#)

Sainsbury, R. M. [68, 76, 77, 84](#)
Sanford, D. [68](#)
Schiffer, Stephen [84, 181](#)
Sellars, W. [141](#)
Skinner, B. F. [11](#)
Skyrms, B. [82](#)
Sloman S. [174](#)
Smith, E. E. [27, 30, 32, 35-7](#)
Stainton, Rob [154](#)
Stich, Steven [28, 174](#)
Strawson, P. F. [138](#)

Tannenbaum, [P.H. 73](#)
Tappenden, J. [69](#)

Turing, A. [55](#)
Tye, Michael [84](#)

van Fraassen, Bas [64-6, 79-82, 84](#)
von Kutschera, F. [68](#)

Watson, John B. [11](#)
Weinreich, U. [103](#)
Williamson, T. [68](#)
Wittgenstein, Ludwig [12, 78, 131](#)
Wright, C. [68](#)

Zalta, Ed [84](#)
Ziff, P. [115](#)

SUBJECT INDEX

References marked with (fn) are found in the footnote of a page. If they're marked (fn) it means that they're not located anywhere else on the page.

- atomic sentences/statements [78](#), [80](#), [90](#), 177
- atomism [25](#), [91](#), [92](#), [93](#), [94](#), [97](#), [111](#), [115](#), [139](#)
- analytic/synthetic (a/s) distinction [14](#), [19–25](#), [40](#), [157](#), [167](#), 174, 189–90, 196, 198
- analyticity [14](#), [18–21](#), [24](#), [34](#), [94](#), [95](#), [97](#), [139](#), [140](#), [142](#) (fn)
- belief [12–13](#), [17–18](#), [40](#), [49](#), [50–2](#), [55](#), [89](#) (fn), [139](#), [149](#), [150–2](#), [156](#), [158](#), 168, 176, 189–90, 198
- Boolean
 - concepts [29–35](#), 170
 - functions 169, 172
- complex expressions *see*
 - primitive/complex expressions
- compositionality [14–25](#), [27–9](#), [31](#), [37–41](#), [43–7](#), [51](#) (compositional structure), [52–62](#), [106](#), [110](#), [112](#), [118–19](#), [139](#), [142–4](#), [146–7](#), 149, 176, 179–82, 199
- compositionality, reverse [34](#)
- compositional semantics *see* semantics, compositional
- concept [26](#) (conceptual networks), [27–37](#), [39–43](#), [48](#) (conceptually), [58–9](#), [60](#) (conceptual systems), [66](#) (conceptual), [71](#) (conceptual), [72](#) (conceptual), [74](#) (conceptual), [79](#) (conceptual), [89](#) (fn), [95](#), [120](#), [135–43](#), 149, [157–9](#), 160 (conceptual framework), [161](#) (conceptually possible), 162–5, [167–73](#), 176, 178 (fn), 185–91, 195–7, 200
 - concept, primitive [32](#), [33](#), [35](#), [137](#), [145](#), 171, 180 (fn), 188
 - conceptual knowledge *see* knowledge, conceptual
 - conceptual role [9](#)
 - conceptual role semantics [140](#)
 - conceptualization [136](#)
 - conceptual truth *see* truth, conceptual
 - constituency, lexical [52](#)
 - constituent [16](#), [18](#), [19](#), [20](#), [22](#), [24](#), [25](#), [28](#), [29](#), [31–5](#), [37](#), [41–3](#), [49](#), [50](#), [59](#), [60](#), [64](#), [67](#), [81](#), [89](#), [111](#), [113](#), [114](#), [118](#), [119](#), [121](#), [126](#), 128, [138](#), [138](#) (fn), [143](#), [144](#), [147](#), 176, 180, 199
 - constituent, conceptual [30](#), [149](#)
 - constituent, lexical [52](#), [96](#), [112](#)
 - constituent, primitive [143](#)
 - constituent, syntactic [14](#), [15](#), [21](#), [52](#)
 - content identity *see* identity of content
 - content similarity *see* similarity of content
- decompositionality *see* lexical decomposition
- definition [22](#), [30](#) (defined in terms of), [44](#), [75](#) (definitional of), [78](#), [85](#), [86](#), [83](#) (undefined), [83](#) (fn), [84](#) (fn), [93](#) (properties defined in terms of . . .), [96](#), [98](#) (fn), [116–18](#), [120](#) (definitional), [121](#) (defining), [127](#), 128 (defining . . .), [144](#) (defined), [145](#) (recursive definitions), [146](#), [150](#) (fn), [157](#), 165, [167](#), 172, 182 (fn), 187, 188 (definitionally), 189 (undefined), 192–3, 196, 197
- denotation [9](#), [66](#), [82–3](#) (fn), [90–2](#), [94–5](#), [98–9](#), [100](#) (fn), 102, [111–16](#), [118](#) (fn), [137](#), [147](#)

- denotational lexicon *see* lexicon, denotational
- empiricism [11](#), [14](#), [70](#), [122](#), [138](#), [156–9](#), [163–5](#), 169–72, 185, 187
- epistemology [10](#), [12](#), [13–14](#), [25](#), [50](#), [95](#) (fn), [143](#), 149, [151](#), 169–70
- extension [15](#), [32](#), [34](#), [40](#), 65–6, [69](#), [74–7](#), [83](#) (fn), [85–6](#), [135](#), [161](#), 178 (fn), 179 (fn), 182
- generative semantics *see* semantics, generative
- generativity [93](#), [98](#) (fn), 102–3, [110–12](#), [115](#), [120](#)
- generativity, lexical [91](#)
- holism [13](#), [22](#), [25](#), [139–40](#) (fn), [142](#) (fn), 164 ('Shopper's Guide' ref.), 171–3, 189 (fn), 192 (fn)
- holism, semantic [23–4](#)
- holophrasism [79](#), [81](#), [82](#) (fn)
- homonymy [103](#), [105](#), [107](#) (fn), [110](#), [116](#), [117](#)
- identity [9](#), [11](#), [13–14](#), [16–19](#), [22–3](#), [26](#), [28](#), [33–4](#), [71](#), [72](#), [74](#), [75](#), [77](#), [80](#) (fn), [82](#) (fn), [91](#), [111](#), [114](#), [138](#), [139](#), [142](#) (fn), [145–7](#), [148](#), [150–1](#), [157](#), [159](#), 162, 170–2, 174, 175, 182, 185, 188, 190–2, 193 (semantic similarity/identity) 194–7
- identity-based meaning theory 177, 186, 187–9
- identity of state spaces 163–4
- identity of content [158](#), [161](#), [164](#), [167–9](#), 171, 175, 178, 180, 182, 183, 198
- idiom [14](#), [39](#), [40](#), [44](#), [56](#), [60](#) (fn), [62](#)
- inference [12](#), [17–25](#), [37](#), [47](#), [49](#), [51–6](#), [61–2](#), [72](#), [89–92](#), [94](#), [96](#), [107–8](#), [130](#), [131](#), [135](#), [138–42](#), [144](#), [147–9](#), [151](#), [153–4](#), [167](#), 170, 172, 188, 194
- inference, analytic &/or compositional inference [18–20](#), [23](#)
- inference, analytic [24](#), [94](#), [96](#), [138](#), [140](#)
- inferential [48](#), [50](#), [54](#), [56](#), [62](#), [89](#), [136](#), [145](#), [156](#), 172, 188
- inferentialism [135](#), [137–40](#), [143–4](#), [147](#), 149–50, [153](#)
- inferential role [13–14](#), [16–25](#), [59](#), [92](#), [136](#), [138–40](#), [141](#) (fn), 142–4, [147](#), 149, [151](#), 187, 189
- inferential roles and meaning *see* meaning and inferential roles
- inferential role semantics (IRS) *see* semantics, inferential role
- knowledge [92](#), [97](#), [152](#)
- knowledge, implicit [52](#), [59](#)
- knowledge, linguistic/conceptual [40](#)
- knowledge, syntactic [101](#)
- knowledge, tacit [49](#)
- lexical decomposition [21](#), [22](#), [26](#), [120](#), [122](#), [124](#)
- lexical generativity *see* generativity, lexical
- lexicalization [121–3](#), [126](#), [130](#)
- lexical meaning *see* meaning, lexical
- lexical primitive [56](#), [57](#), [127](#), [130–1](#)
- lexicon [20–2](#), [25–6](#), [36–7](#), [43](#), [45–7](#), [51](#), [56](#), [58–61](#), [62](#) (fn), [69](#), [73](#), [89–119](#), [121–3](#), [127](#), [130–1](#), 181 (fn), 187
- lexicon, denotational [93](#), [97](#)
- lexicon, sense enumeration 103–4, [112](#), [116–17](#)
- meaning [9](#), [11–12](#), [15](#), [17](#), [21](#), [36](#) (fn), [37](#), [38](#), [39](#), [43](#), [44](#), [45](#) (theories of), [46](#), [49](#), [51](#), [52](#), [53](#), [57](#), [58](#), [59](#), [60](#), [61](#) (complex), [62](#) (fn), [63](#) (fn), [64](#), [69](#), [70](#), [72](#), [74–6](#), [79](#), [81–2](#), [83](#) (fn), [86](#), [89–91](#), [91](#) (theories of), [92](#), [93](#), [94](#), [95](#) (fn), [97–9](#), 103–4, [106–7](#), [109–13](#), [115–16](#), [120](#) (fn), [121](#) (fn), [122](#), [127](#) (well-formed), [127–8](#), [131](#), [138](#), [139](#), [140](#), [158](#), 164, 172 (meaning holism), 176, 180, 183, 185, 187, 198
- meaning, compositionality of [18](#), [44](#), [47](#), [56](#)
- meaning, identity of [24](#), [155](#), 164, 168, 179 (fn), 199–200
- meaning, inferential roles and [13](#), [14](#), [16](#), [18–25](#)
- meaning, lexical [21–2](#), [25](#), [28](#), [43](#), [45](#), [57–62](#), [69–70](#), [91](#), [97–9](#), 102–3, [112](#)
- meaning, similarity of 24–5, [155](#), 168, 174–5, 179, 181–2, 199–200

- meaning, theory of [9](#), [10](#), [45](#), [91](#), [138](#),
[149](#), [153](#), [174](#), [176](#), [177](#), [178](#), [186](#),
[187](#), [188](#), [189](#)
identity-based *see* identity-based
meaning theory
similarity-based *see* similarity-based
meaning theory
- metaphysics [30](#) (fn), [44–5](#), [48–51](#), [89](#),
[98](#), [105](#), [139](#) (fn), [144](#), [149](#), [162](#),
[169](#), [174–5](#), [187](#)
- necessity [38](#), [72–3](#), [78](#), [80–1](#), [85–6](#), [91](#),
[94](#), [96](#), [97](#), [98](#) (fn), [99](#), [103](#), [113](#)
(fn), [149](#), [153](#), [162](#), [169](#)
- necessity, conceptual [69](#), [70](#), [75–6](#), [81–6](#)
- necessary condition [42](#), [77](#), [95](#), [101](#),
[139](#), [163](#)
- necessary truth *see* truth, necessary
- noun [16](#), [92](#) (fn), [98](#), [104–5](#), [116–17](#),
[120](#), [128](#), [146](#)
- ontology [13–14](#), [50](#), [105](#), [107](#), [109](#) (fn),
[176](#) (fn), [199](#)
- polysemy [103–5](#), [107](#) (fn), [109–11](#), [112](#)
(fn), [113](#), [115–17](#), [118](#) (fn), [119](#)
- precisification [66](#), [70–1](#), [73–7](#), [81](#), [85–6](#)
- primitive/complex expressions [47](#),
[56–7](#), [59](#)
- primitive concept *see* concept, primitive
- primitive words [126](#), [127](#), [129–31](#)
- productivity [15](#), [24](#), [27–31](#), [33–4](#), [61](#),
[117](#), [142–3](#), [154](#), [176](#), [181](#) (fn), [182](#)
(fn)
- prototype [27–42](#), [59](#), [104](#), [170](#), [172](#),
[186](#), [188](#), [200](#)
- prototype theory [28–30](#), [32](#), [34–45](#)
- recalibration [37–9](#)
- relativism [13–14](#), [22–3](#), [31](#), [105](#), [122](#)
(fn), [145](#), [152](#), [179](#), [185–7](#),
[189–92](#)
- reliabilism [36](#), [141–42](#), [149](#), [151–2](#)
- reverse compositionality *see*
compositionality, reverse
- semanticality [97–8](#)
- semantics [9–11](#), [13–26](#), [29](#), [32–3](#), [37](#),
[39](#), [56–61](#), [63–5](#), [67–8](#), [70–1](#), [73–4](#),
[77–82](#), [84](#), [86](#), [92–3](#), [95–6](#), [100–2](#),
[104](#) (sense enumeration semantic
analysis), [105](#), [106](#) (fn), [107](#),
[109–10](#), [111–12](#), [116](#), [118](#), [120](#),
[135](#), [136–40](#), [141](#) (fn), [142](#) (fn),
[143–4](#), [147](#), [149–51](#), [154–6](#), [159](#),
[161](#), [165–70](#), [172](#), [177](#), [179](#), [181](#),
[189–93](#), [195–7](#), [198](#), [199](#)
- semantics
compositional [39](#), [57](#), [113](#)
denotational [110](#)
generative [26](#), [102](#), [126](#), [130](#)
inferential role (IRS) [13](#), [89–91](#), [93](#),
[97](#), [136](#), [189](#)
interpretive [26](#), [130](#)
network [157–8](#)
referential [117](#)
state-space [155–6](#), [158–73](#), [196](#)
vector-space [170](#), [184](#), [185](#), [186](#), [190](#)
- semantic property [79](#), [99](#), [180](#), [188](#),
[194](#), [196](#)
- semantic similarity *see* similarity,
semantic
- semantic space [159–67](#), [170](#), [174](#), [185](#)
(fn), [186](#), [188–91](#), [193–7](#), [199](#)
- semantic theory [12](#), [45](#), [93](#), [97](#), [150](#),
[156](#), [157](#), [159](#), [165–6](#), [174–7](#), [181](#),
[183](#)
- semantic value [33](#), [65](#), [81](#), [113](#)
- semantic well-formedness [97–9](#)
- semantic interpretation [16](#), [39](#), [113–15](#)
- semantic representation [99](#), [111](#), [121–2](#)
- sense enumeration lexicon *see* lexicon,
sense enumeration
- similarity [30](#), [34–5](#), [100](#), [142](#) (fn), [155](#),
[158](#), [160](#), [162–3](#), [166–7](#), [168](#)
(conceptual), [178–81](#), [182](#)
(similarity-based translation
theory), [183](#), [185](#), [187](#), [189–94](#),
[196](#), [198](#)
- similarity-based meaning theory [177](#),
[184](#)
- similarity of content [157–9](#), [161–2](#), [164](#),
[168](#), [170–1](#), [173–5](#), [180](#), [183–4](#),
[186](#), [188](#), [194](#) (content-state
similarity), [197–8](#)
- similarity, semantic [156](#), [163](#), [164](#), [169](#),
[178–9](#), [183](#), [193–4](#), [198](#)
- stereotype [27](#), [32](#), [43](#), [58](#), [59–60](#), [62](#),
[199–200](#)
- subjective probabilities [17](#), [18](#)

- supervaluation [31](#), [32](#) (fn), [63–72](#), [74](#), [78–81](#), [83](#), [84](#) (fn), [86](#)
- synonymy [10](#), [26](#), [93–7](#), [121](#), [123](#), [126](#) (fn), [127](#), [174–5](#), [182](#)
- syntactic constituents [14](#), [15](#), [21](#), [52](#), [179](#)
- syntactic knowledge *see* knowledge, syntactic
- syntactic structure [14](#), [16](#), [81](#), [82](#), [97](#), [116](#)
- syntax [15–16](#), [25–6](#), [29](#) (fn), [45–7](#), [49](#), [51](#), [55](#), [57–9](#), [60](#) (fn—syntactic conjunct), [62](#) (fn), [65](#), [73](#), [81](#), [99–103](#), [113–14](#), [116](#), [121](#), [127](#), [128](#) (fn), [131](#), [140](#), [143–4](#), [146–7](#), [179](#), [180](#) (fn), [181](#) (fn), [182](#) (fn)
- systematicity [15](#), [24](#), [35](#), [58](#), [60–1](#), [99](#), [142](#) (fn), [143](#) (fn), [161](#), [176](#)
- truth, conceptual [68–73](#), [77](#), [85](#)
- truth conditions [15](#), [25](#), [31](#), [32](#), [67](#), [73](#), [75](#), [82](#), [153](#), [176](#), [177](#)
- truth, definite [32](#), [63](#) (fn), [65–7](#), [71–2](#), [74–9](#), [81](#)
- truth, necessary [60](#) (fn), [75](#), [115](#)
- truth value [29](#), [30](#) (fn), [31](#), [32](#) (fn), [63–6](#), [72](#), [75](#), [79–86](#), [96](#), [179](#)
- truth value (tv) gaps [29–31](#), [32](#) (fn), [63–5](#), [67–8](#), [80](#), [83](#), [86](#)
- vagueness [14](#), [27–32](#), [63–82](#), [84](#), [86](#)
- verb [92](#), [97](#), [99](#), [101–2](#), [104–5](#), [107–9](#), [111–12](#), [115](#), [117](#), [120–4](#), [126](#), [128–9](#)



Jerry Fodor and Ernie Lepore have produced a series of original and controversial essays on issues relating to compositionality in language and mind; they have now revised them all for publication together in this volume. Compositionality is the following aspect of a system of representation: the complex symbols in the system inherit their syntactic and semantic properties from the primitive symbols of the system. Fodor and Lepore argue that compositionality determines what view we must take of the nature of concepts. Anyone trying to figure out how language and mind work must take account of this challenging work by two leading figures in the field.

Jerry A. Fodor is Professor of Philosophy at Rutgers University, New Jersey.

Ernie Lepore is Director of Rutgers University Center for Cognitive Science.

OXFORD
UNIVERSITY PRESS

www.oup.com

Cover illustration:
Getty images © PhotoDisc

ISBN 0-19-925216-5



9 780199 252169

Copyrighted material